A Framework to Identify

Epigenome and Transcription Factor Crosstalk

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät der

Universität Basel

von

Phil Pascal Arnold

aus

Basel, Schweiz

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Erik van Nimwegen

Prof. Dr. Dirk Schübeler

Basel, 20. September 2011

Prof. Dr. Martin Spiess

Dekan

creative commons

Attribution-Noncommercial-No Derivative Works 2.5 Switzerland

---

You are free:

to Share — to copy, distribute and transmit the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license) available in German:
http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de

Disclaimer:
The Commons Deed is not a license. It is simply a handy reference for understanding the Legal Code (the full license) — it is a human-readable expression of some of its key terms. Think of it as the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its contents do not appear in the actual license. Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying of, or linking to this Commons Deed does not create an attorney-client relationship.

Quelle: http://creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en Datum: 3.4.2009

This degree is dedicated to
my mother and father Ruth and Peter,
my brother Mirko,
and my two little sisters
Céline and Sophia.

'Strength and honor.'
[Maximus - Gladiator]

Acknowledgments

The successful completion of this thesis would not have been possible without the assistance and support of many people.

First of all, I would like to thank Erik van Nimwegen, who offered me the opportunity to undertake the work towards this degree. His guidance and advice throughout the years have truly been essential to finally reach this goal.

I am very grateful to Anne Schöller and Dirk Schübeler for a textbook example of a collaboration, useful discussions, and fruitful & entertaining meetings.

I am especially indebted to Yvonne Steger for taking care of all the administrative things during these last 4 years, thereby making my life so much easier. Also, special thanks to the IT-guys Konstantin, Rainer, and Jan for cleaning up all the computer mess I made.

I would like to thank Piotr Balwierz, Nacho Molina, Lukas Burger, Mikhail Pachkov, Evgeniy Ozonov, Luise Wolf, Florian Geier, Peter Pemberton-Ross, and Nick Kelley for invaluable help with many problems I encountered in my work.

Also, I am much obligated to Nick Kelley and Peter Pemberton-Ross for proofreading my manuscripts and eliminating lots of mistakes.

And finally, I would like to that Erik van Nimwegen, Drik Schübeler, and Renato Paro for serving on my examining committee.

To everyone I mentioned and those I forgot: **Thank you very much!**

Introduction

After a short introduction on the biology of chromatin and Epigenetics, especially Polycomb, I am going to describe our work on a mathematical framework to predict crosstalk between transcription factors and epigenetic marks. In particular, the first part of the thesis focuses on the application of Epi-MARA, an algorithm to model relative changes of chromatin levels at different cell stages as a linear function of the expected number of predicted transcription factor binding sites, to an in-vitro mouse neuronal differentiation system. We investigate on the genome-wide dynamics of the try-methylation of lysine 27 on histone 3. One of the predicted transcription factors, REST, is intensively studied and its predicted properties are verified by experiments.

The second part of the thesis describes in detail our frequently used transcription factor binding site prediction algorithm, called MotEvo, and the algorithm to model relative changes of chromatin or expression levels at several time points or different tissues as a linear function of the expected number of predicted transcription factor binding sites. This algorithm comes in two versions and is called MARA when applied to expression data and Epi-MARA when applied to epigenetic marks.

The discussion will focus on the main findings of my PhD thesis and give an outlook of where future work could be taken up.

Contents

I Epigenetic Motif Activity Response Analysis: A Framework to Identify Epigenome and Transcription Factor Crosstalk 1

1 Introduction 3
1.1 Introduction . . . 3
1.1.1 Chromatin Modifications and Polycomb . . . 3
1.1.2 Recruitment of chromatin modifications and Polycomb . . . 5
1.1.3 Epi-MARA . . . 6

2 Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting 13
2.1 Introduction . . . 14
2.2 Results . . . 15
2.2.1 Predicting mediators of chromatin changes using Epi-MARA . . . 15
2.2.2 Experimentally determined REST binding sites support the computational prediction 17
2.2.3 REST binding is associated with H3K27me3 dynamics genome-wide . . . 18
2.2.4 REST protein is required for local H3K27 methylation levels . . . 20
2.2.5 REST affects H3K27me3 and expression independently at many target genes . . . 20
2.2.6 Promoter fragments containing REST or SNAIL binding sites locally recruit methylation of H3K27 . . . 20
2.3 Discussion . . . 22
2.4 Acknowledgements . . . 25
2.5 Methods . . . 27
2.5.1 Epi-MARA . . . 27
2.5.2 Cell Culture . . . 28
2.5.3 Western Blot Analysis . . . 28
2.5.4 Immunocytochemistry . . . 28
2.5.5 Chromatin-IP . . . 28
2.5.6 Quantitative real time PCR . . . 28
2.5.7 Next generation sequencing . . . 28
2.5.8 Genomic coordinates . . . 29
2.5.9 Read filtering, alignment and weighting . . . 29
2.5.10 Analysis of sequencing data . . . 29
2.5.11 RNA preparation and expression analysis . . . 31
2.5.12 Recombinase mediated cassette exchange (RMCE) . . . 31
2.6 Bibliography . . . 32
2.7 Supplementary Figures . . . 37
2.8 Supplementary Tables . . . 48

2.9 Supplementary Methods . . . 50
2.9.1 Epi-MARA . . . 50

3 MotEvo: Integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences 55
3.1 Introduction . . . 55
3.2 Methods . . . 56
3.2.1 Binding site configurations . . . 57
3.2.2 Probabilities under the evolutionary model . . . 57
3.2.3 Unidentified Functional Elements . . . 59
3.2.4 Forward/backward algorithm . . . 60
3.2.5 Transcription Factor Binding Site Predictions . . . 61
3.2.6 Prior Updating . . . 62
3.2.7 Enhancer prediction . . . 62
3.2.8 Weight Matrix Refinement . . . 62
3.3 Results . . . 63
3.3.1 The UFE model strongly reduces spurious predictions . . . 63
3.3.2 MotEvo's novel features improve TFBS prediction . . . 63
3.3.3 WM refinement improves TFBS predictions . . . 65
3.3.4 Enhancer prediction accuracy increases with the number of species used . . . 66
3.4 Discussion . . . 67
3.5 Supplementary Methods . . . 69
3.5.1 Likelihood of an alignment column . . . 69
3.5.2 Higher order background models within the evolutionary model . . . 69
3.5.3 Calculating the probabilities under the UFE model . . . 70
3.5.4 Backward recursion relation . . . 71
3.5.5 Weight matrix refinement . . . 71
3.6 Construction of benchmarking regions . . . 73
3.7 The UFE model strongly reduces spurious predictions . . . 75
3.8 Species selection improves TFBS predictions . . . 76
3.9 WM refinement . . . 77
3.9.1 Refining motifs outperforms ordinary motif inference . . . 77
3.10 Enhancer prediction . . . 82
3.11 Comparison to MONKEY and PhyloScan . . . 84
3.12 Dependence of MotEvo's TFBS predictions on different aligners . . . 86

4 ISMARA: Modeling genomic signals as a democracy of regulatory motifs 91
4.1 Introduction . . . 91
4.2 Results . . . 93
4.2.1 An Integrated System for Motif Activity Response Analysis . . . 93
4.2.2 Overview of the results presented by ISMARA . . . 95
4.2.3 Inferring motif activity dynamics: inflammatory response . . . 98
4.2.4 Identifying novel master regulators: Mucociliary differentiation of bronchial epithelial cells . . . 99
4.2.5 Epithelial-Mesenchymal Transition: including microRNAs in core regulatory networks 101
4.2.6 TF activities effecting chromatin state: analysis of ChIP-seq data . . . 102
4.3 Discussion . . . 105
4.4 Methods . . . 107
4.4.1 Materials . . . 108

4.5 Supplementary Methods . . . 109
4.5.1 Human and mouse promoteromes . . . 109
4.5.2 A curated set of regulatory motifs . . . 110
4.5.3 Transcription factor binding site predictions . . . 112
4.5.4 Associating miRNA target sites with each promoter . . . 113
4.5.5 Expression data processing . . . 114
4.5.6 ChIP-seq data processing . . . 115
4.5.7 Motif activity fitting. . . . 116
4.5.8 Processing of replicates . . . 118
4.5.9 Target predictions . . . 119
4.5.10 Principal component analysis of the activities explaining chromatin mark levels . . . 121
4.6 Fraction of variance explained by the fit . . . 124
4.7 Overview of results presented in the web-interface . . . 125
4.8 HNF1a activity in pancreas . . . 133
4.9 Reproducibility of motif activities . . . 133
4.10 Motifs dis-regulated in tumor cells . . . 135
4.11 XBP1 motif activity and mRNA expression . . . 135
4.12 Analysis of the ENCODE ChIP-seq data . . . 135
4.12.1 PCA analysis . . . 139

5 Discussion and Future Work **153**
5.0.2 Biological relevance . . . 154
5.0.3 Future work . . . 155

Part I

Epigenetic Motif Activity Response Analysis: A Framework to Identify Epigenome and Transcription Factor Crosstalk

Chapter 1

Introduction

1.1 Introduction

Multi-cellular life, to which we all belong, is possible through the coordination of genetically identical cells, each performing their specific sets of tasks. Although they share the same genome, life has evolved the ability to control which genes are active through a series of cell fate decisions, starting at the embryo, called differentiation. These expression patterns, in turn, define the tissues and cellular phenotypes which comprise the organism. Further more, mechanisms exists which stabilize the expression states of these cells, and which allow the inheritance of these states from one generation to the next following cell division.
Histone modifications, histone variant composition, nucleosome positioning, proteins that bind directly to DNA or to histone modifications, higher order chromatin structure, and non-coding RNA [1–5] are several ways of carrying the (local) chromatin structure and, therefore, using the genome in different ways (field of Epigenetics, see figure 1.1).

One of the current topics in the field of Epigenetics is how chromatin modifications are targeted to their sites of action. In our work, we have focused on predicting transcription factors (TFs) that are involved in recruiting or depleting histone modifications, accompanied by rigorous experimental validation.

1.1.1 Chromatin Modifications and Polycomb

There are a number of chromatin modifications, namely acetylation, methylation, phosphorilation, and ubiquitylation [1]. Even though our approach of identifying key TFs involved in recruitment of chromatin modifications can be applied to any kind of chromatin modifications, our main analysis has focused on the methylation of histones.
Mono-, di-, or tri-methylation of histones can occur either at lysine or arginine residues. In the case of lysine, the methylation pattern is set by lysine methyltransferases [6]. Its counterpart is a lysine-specific demethylase (LSD1 enzyme), which can only remove mono- and di-methylation marks [7]. Another class of a histone demethylase, Jumonji histone demethylase, is likely to contain enzymes that are able to remove tri-methylation marks [7]. Arginine methylation is mediated by arginine methyltransferases and is removed by PADI4 [7, 8]. In most cases, the effect of chromatin modifications on gene regulation and the chromosomal structure is unknown. However, ChIP-seq experiments revealed that methylation of lysine 4 on histone 3 (H3K4), H3K36, and H3K79 are mainly found at promoters of active genes, and, therefore, these marks are thought to be involved in transcription [9].

One of the best studied chromatin modification is tri-methylation of lysine 27 on histone 3 (H3K27me3) in

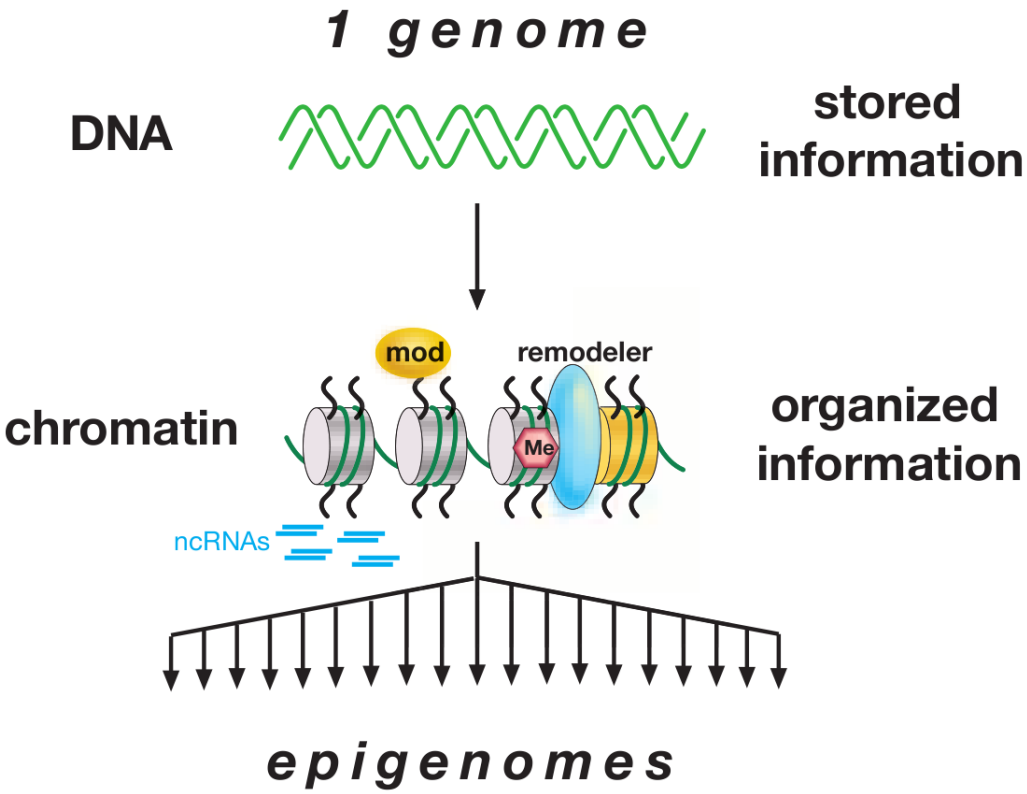


Figure 1.1: Epigenomes (Adapted from [1]): All cells in an organism have the same genome, but different chromatin compositions lead to different gene expression profiles and, eventually, different cell shapes and functions. The chromatin composition (epigenome) includes histone modifications, histone variants, DNA binding proteins, nucleosome positioning and more.

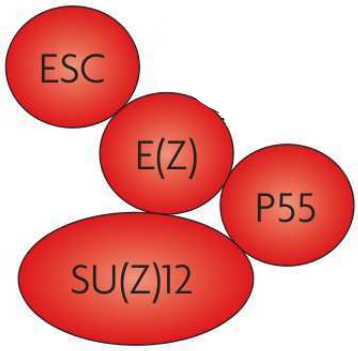


Figure 1.2: PRC2 (Adapted from [12]): The core of the PRC2 in drosophila melanogaster consists of four proteins depicted in red [11]. ESC supports protein-protein interactions with E(Z) and p55. E(Z) contains a SET domain that posseses histone lysine methylation activity [1]. The protein p55 interacts with ESC. SU(Z)12 interacts with ESC [3]. The PRC2 complex is highly conserved in invertebrates, vertebrates, and plants [1].

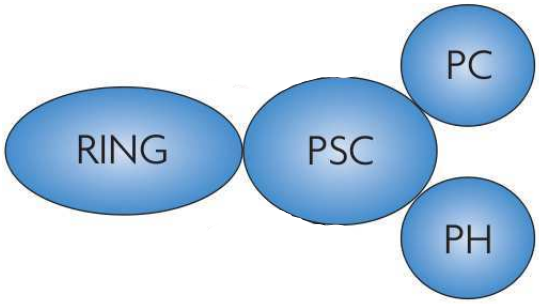


Figure 1.3: PRC1 (Adapted from [12]): The PRC1 complex in drosophila melanogaster consists of four core proteins depicted in blue [14]. First of all, there is the Polycomb protein (PC). PC possesses a chromodomain that binds to H3K27 and H3K9 [15]. PH contains a SAM domain involved in protein-protein interactions. These protein-protein interactions could be used by PCR1 to form large nuclear complexes required for silencing [1, 16]. PSC might be a mediator of protein-protein interactions [1]. RING is necessary to maintain ubiquitylated H2A on the inactive X chromosome [17, 18].

drosophila melanogaster. H3K27me3 is mediated by Polycomb Group proteins and is involved (amongst others) in controlling HOX gene activity throughout the development of the fly, long after the initial TFs have disappeared [1]. This is achieved by remodeling the local chromatin structure and by interfering with RNA Polymerase II and, thereby, epigenetically - involving a modification in gene expression that is independent of the DNA sequence of a gene - silencing genes [10, 11]. Two key players are thought to cooperate for Polycomb silencing, namely Polycomb Repressive Complex 1 (PRC1) and Polycomb Repressive Complex 2 (PRC2) [12] (see figure 1.3). PRC2 has histone methyltransferase activity and, therefore, sets the H3K37me3 mark at regions labeled for silencing beforehand. In a second step, it is assumed that PRC1 binds to the H3K27me3 mark and, thereby, changes the structure of the chromatin leading to gene repression [1, 13]. While it is often the case that both PRC2 and PRC1 are required for gene silencing, certain genes are targeted by only one or the other [13].

1.1.2 Recruitment of chromatin modifications and Polycomb

After this short overview on chromatin modifications and their functions, we are going to focus on how these chromatin modifications, especially Polycomb, are thought to be recruited. For Polycomb, experiments in drosophila melanogaster showed that PRC2 is targeted by DNA sequences, so called Polycomb response elements (PREs) [1,12,13]. The length of these PREs vary from a few hundred to several thousand basepairs [1]. A careful examination of the PREs' sequence composition revealed a handful of DNA binding TFs that seem to determine PRE function. For instance, the zinc finger protein Pleiohomeotic (PHO), GAGA factor (GAF), specificity protein 1 (SP1), and a few more [19–22]. Genome-wide studies showed the central role of PHO in PRE function, whereas the other factors do not seem to be of such importance [3].

In mammals, only a few Polycomb response like elements have been currently defined [23,24]. The main difficulty is that the binding of PRC1 and PRC2 is spread across many kilobases of mammalian developmental

genes [25–27], making it very hard to confidently identify PREs. A promising candidate is ying-yang (YY1), the homologue of drosophilas PHO [28]. However, the overlap of YY1 and PRC2 in, for instance, embryonic stem cells is not very pronounced [29]. Therefore, YY1 is unlikely to be one of the main Polycomb recruiters. Another approach has been to perform ChIP experiments followed by deep sequencing (ChIP-seq) to track down the genome-wide occupancy PRC1 and PRC2. Surprisingly, PRC2 almost always overlaps with CpG islands or CpG enriched regions. This suggests that CG-binding proteins are involved in Polycomb recruitment. However, no sequence elements have as of yet been identified [27]. There are other examples of how chromatin modifications, especially H3K27, are recruited. By studying models for epigenetic silencing like the inactivation of the X-chromosome and the HOX gene clusters, long non-coding RNAs and H3K27 can be found. For HOTAIR, a 2.2kbs long non-coding RNA involved in tri-methylation of the HOXD locus [30], an interaction with a PRC2 unit has experimentally been determined. Therefore, it is hypothesized that the Polycomb complex can also be recruited by RNA motifs [31].

1.1.3 Epi-MARA

To elaborate on the idea that chromatin modifications are at least partially recruited by TFs, we have developed a systematic computational approach, called Epi-MARA, using genome-wide chromatin modification measurements to identify key TFs ab initio. Our approach builds on the Motif Activity Response Analysis (MARA) that was developed previously [32].

The main idea is to use measurements of a given chromatin mark at different conditions or time-points as an input and model their dynamic as a linear function of predicted transcription factor binding sites (TFBSs) in regulatory regions genome-wide:

$$M_{pt} = \text{noise} + c_p + \sum_m N_{pm} A_{mt}, \tag{1.1}$$

where c_p is the basal level of the chromatin mark at region p, M_{pt} is the chromatin modification at region p at time t, N_{pm} is the expected number of binding sites at this region p for motif m, and A_{mt} is the inferred recruitment or depletion activity of motif m at time t. Whenever EpiMARA infers a highly positive activity A_{mt}, this predicts that the binding TF recruits the chromatin mark, whereas a highly negative A_{mt} implies that the binding TF inhibits deposition of the mark. It is worth mentioning that the term c_p gives rise to a row normalization of the matrix M_{pt}, so at the end of the day, we model the dynamics of the chromatin modifications in terms of predicted TFBSs.
While the chromatin marks can 'easily' be measured using generally accepted experimental techniques like ChIP-chip and ChIP-seq, obtaining TFBSs for all known TFs for a set of regulatory regions is much a bigger challenge. Obviously, performing a ChIP-seq measurement for each TF and each condition or time point is not (yet) an option. Thus, instead of experimentally measuring the binding sites of the different factors, we computationally predict them. There are number of ways how this can be done. In our case, we inferred a set of ≈ 200 positional weight matrices (WMs) from experimentally validated binding sites. Each WM represents a DNA binding pattern (motif) for one or several TFs. Then, for all regulatory regions of interest, for instance all promoters, we constructed multiply-aligned regions. Finally, we run the MotEvo [33] algorithm that takes as input a WM and multiply-aligned regions and calculates for each promoter p and each motif m the expected number of TFBSs. We then summarize all the information in a matrix N_{pm}. Using standard numerical procedures (for instance singular value decomposition), and a Gaussian noise term, the resulting system of equations can now be solved to infer the recruitment or depletion activity for each of the TFs (for more details see chapter 4).

A collaboration with Dirk Schübeler's lab led to the first biological application of Epi-MARA. Having an in vitro neuronal differentiation system at hand, in which mouse embryonic stem (ES) cells differentiate through

a neural progenitor (NP) stage into terminally differentiated neurons (TN), we first focused on the chromatin dynamics of the repressive chromatin modification H3K27me3 at promoters. Our initial Epi-MARA analysis predicted the TF REST as a key factor that recruits H3K27me3 from the ES to the NP stage. At that time, REST was known to repress neuronal genes in non-neuronal tissues [34], but no association with Polycomb was known.
Next, we identified all regions in the genome that are clearly enriched in H3K27me3. By looking at the CpG content of these H3K27me3 clusters, we could divide them into high-CpG ($\approx$ 7700) and low-CpG ($\approx$ 10500) H3K27me3 clusters. Using ChIP-seq binding data for REST in ES and NP allowed us to identify each H3K27me3 cluster as a REST target or as a non-target. ChIP-seq for TN was not performed because REST is not expressed at that stage. It is worth mentioning that most of the H3K27me3 clusters that lie within 2kb from a known transcription start site (TSS) belong to the high-CpG class. Finally, running Epi-MARA on all H3K27me3 clusters revealed some striking differences between CpG-high and CpG-low H3K27me3 REST targets: high-CpG REST targets seem to gain H3K27me3 from ES to NP, whereas low-CpG REST targets seem to lose H3K27me3 from ES to NP (see figure 2.3b in chapter 2). Looking at the overall fold-change distribution of the H3K27me3 levels across the three times points ES, NP, and TN, we realized this different behavior of CpG-high and CpG-low H3K27me3 regions occurs to all H3K27me3 enriched regions, though less marked than at the REST targets (see figure 2.3c, d in chapter 2).

The recruitment of H3K27me3 by REST at high-CpG regions and the depletion of H3K27me3 by REST at low-CpG regions at the NP stage was validated in two independent ways: First, by comparing H3K27me3 levels between wild-type and REST ko cells clearly showed that in NP high-CpG REST targets have more H3K27me3 in the wild-type. On the other hand, low-CpG REST targets clearly have less H3K27me3 in the wild-type. For both high-CpG and low-CpG REST targets, the difference in ES is much smaller (see figure 2.4b in chapter 2). Second, the fold-change plots mentioned before clearly show what Epi-MARA predicted. In the case of the high-CpG regions, we even had a third experimental validation. By inserting promoter fragments containing wild-type or mutated REST binding sites into H3K27me3 free regions in the mouse genome, we could show that the REST binding site is indeed necessary to recruit H3K27me3.

Together our results firmly establish REST as an important recruiter of Polycomb repression during early neurogenesis.

Bibliography

[1] C. David Allis, Thomas Jenuwein, and Danny Reinberg. *Epigenetics*. Cold Spring Harbor Laboratory Press, 2007.

[2] Fabio Mohn. *Epigenome plasticity during cellular differentiation*. PhD thesis, FMI, 2009.

[3] Jeffrey A. Simon and Robert E. Kingston. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Biol*, 10:697–708, 2009.

[4] Aline V Probst, Elaine Dunleavy, and Genevieve Almouzni. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol*, 10:192–206, 2009.

[5] James Flanagan and Laurence Wild. An epigenetic role for noncoding rnas and intragenic dna methylation. *Genome Biology*, 8(6):307, 2007.

[6] Monika Lachner, Donal O'Carroll, Stephen Rea, Karl Mechtler, and Thomas Jenuwein. Methylation of histone h3 lysine 9 creates a binding site for hp1 proteins. *Nature*, 410:116–120, 2001.

[7] Robert J. Klose, Eric M. Kallin, and Yi Zhang. Jmjc-domain-containing proteins and histone demethylation. *Nat Rev Genet*, 7:715–727, 2006.

[8] J. Doyne Farmer. Physicists attempt to scale the ivory towers of finance. *Computing in Science and Engineering*, 1999.

[9] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823 – 837, 2007.

[10] Valerio Orlando. Polycomb, epigenomes, and control of cell identity. *Cell*, 112(5):599 – 606, 2003.

[11] Stuart S. Levine, Ian F.G. King, and Robert E. Kingston. Division of labor in polycomb group repression. *Trends in Biochemical Sciences*, 29(9):478 – 485, 2004.

[12] Yuri B. Schwartz and Vincenzo Pirrotta. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet*, 8:9–22, 2007.

[13] Raphael Margueron and Danny Reinberg. The polycomb complex prc2 and its mark in life. *Nature*, 469:343 – 349, 2011.

[14] Nicole J. Francis and Robert E. Kingston. Mechanisms of transcriptional memory. *Nat Rev Mol Cell Biol*, 2(6):409–421, 2001.

[15] Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD, and Khorasanizadeh S. Molecular basis for the discrimination of repressive methyl-lysine marks in histone h3 by polycomb and hp1 chromodomains. *Genes Dev*, 17(15):1870–1881, 2003.

[16] Andrew J. Saurin, Carol Shiels, Jill Williamson, David P.E. Satijn, Arie P. Otte, Denise Sheer, and Paul S. Freemont. The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. *J. Cell. Biol.*, 142:887–898, 1998.

[17] Jia Fang, Taiping Chen, Brian Chadwick, En Li, and Yi Zhang. Ring1b-mediated h2a ubiquitination associates with inactive x chromosomes and is involved in initiation of x inactivation. *Journal of Biological Chemistry*, 279(51):52812–52815, 2004.

[18] Ru Cao, Yu ichi Tsukada, and Yi Zhang. Role of bmi-1 and ring1a in h2a ubiquitylation and hox gene silencing. *Molecular Cell*, 20(6):845 – 854, 2005.

[19] Bernd Schuettengruber, Daniel Chourrout, Michel Vervoort, Benjamin Leblanc, and Giacomo Cavalli. Genome regulation by polycomb and trithorax proteins. *Cell Press*, 128:735–745, 2007.

[20] Jurg Muller and Judith A Kassis. Polycomb response elements and targeting of polycomb group proteins in drosophila. *Current Opinion in Genetics & Development*, 16(5):476 – 484, 2006. Differentiation and gene regulation.

[21] Renato Ringrose and Renato Paro. Polycomb/trithorax response elements and epigenetic memory of cell identity. *Development*, 134:223–232, 2007.

[22] Ru Cao, Liangjun Wang, Hengbin Wang, Li Xia, Hediye Erdjument-Bromage, Paul Tempst, Richard S. Jones, and Yi Zhang. Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. *Science*, 298(5595):1039–1043, 2002.

[23] Angela Sing, Dylan Pannell, Angelo Karaiskakis, Kendra Sturgeon, Malek Djabali, James Ellis, Howard D. Lipshitz, and Sabine P. Cordes. A vertebrate polycomb response element governs segmentation of the posterior hindbrain. *Cell*, 138:885–897, 2009.

[24] Caroline J. Woo, Peter V. Kharchenko, Laurence Daheron, Peter J. Park, and Robert E. Kingston. A region of the human hoxd cluster that confers polycomb-group responsiveness. *Cell*, 140:99, 110.

[25] Laurie A. Boyer, Kathrin Plath, Julia Zeitlinger, Tobias Brambrink, Lea A. Medeiros, Tong Ihn Lee, Stuart S. Levine, Marius Wernig, Adriana Tajonar, Mridula K. Ray, George W. Bell, Arie P. Otte, Miguel Vidal, David K. Gifford, Richard A. Young, and Rudolf Jaenisch. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441:349, 353.

[26] Tong Ihn Lee, Richard G. Jenner, Laurie A. Boyer, Matthew G. Guenther, Stuart S. Levine, Roshan M. Kumar, Brett Chevalier, Sarah E. Johnstone, Megan F. Cole, Kyo ichi Isono, Haruhiko Koseki, Takuya Fuchikami, Kuniya Abe, Heather L. Murray, Jacob P. Zucker, Bingbing Yuan, George W. Bell, Elizabeth Herbolsheimer, Nancy M. Hannett, Kaiming Sun, Duncan T. Odom, Arie P. Otte, Thomas L. Volkert, David P. Bartel, Douglas A. Melton, David K. Gifford, Rudolf Jaenisch, and Richard A. Young. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301 – 313, 2006.

[27] Manching Ku, Richard P. Koche, Esther Rheinbay, Eric M. Mendenhall, Mitsuhiro Endoh, Tarjei S. Mikkelsen, Aviva Presser, Chad Nusbaum, Xiaohui Xie, Andrew S. Chi, Mazhar Adli, Simon Kasif, Leon M. Ptaszek, Chad A. Cowan, Eric S. Lander, Haruhiko Koseki, and Bradley E. Bernstein. Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. *PLoS Genet*, 4(10):e1000242, 10 2008.

[28] Matthew J. Thomas and Edward Seto. Unlocking the mechanisms of transcription factor yy1: are chromatin modifying enzymes the key? *Gene*, 236(2):197 – 208, 1999.

[29] Sharon L. Squazzo1, Henriette OGeen1 andn Vitalina M. Komashko, Sheryl R. Krig, Victor X. Jin, Sung wook Jang, Raphael Margueron, Danny Reinberg, Roland Green, and Peggy J. Farnham. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Research*, 16:890–900, 2006.

[30] John L. Rinn, Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha A. Brugmann, L. Henry Goodnough, Jill A. Helms, Peggy J. Farnham, Eran Segal, and Howard Y. Chang. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *CELL*, 129:1311–1323, 2007.

[31] Jing Zhao, Bryan K. Sun, Jennifer A. Erwin, Ji-Joon Song, and Jeannie T. Lee. Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome. *Science*, 322(5902):750–756, 2008.

[32] The FANTOM Consortium and Riken Omics Science Center. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, 41:553–562, 2009.

[33] Erik van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6:S4, 2007.

[34] CJ Schoenherr and DJ Anderson. The neuron-restrictive silencer factor (nrsf): a coordinate repressor of multiple neuron-specific genes. *Science*, 267(5202):1360–1363, 1995.

Chapter 2

Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting

Phil Arnold[1], Anne Schöler[1], Mikhail Pachkov, Piotr Balwierz, Helle Jorgensen, Michael B. Stadler, Erik van Nimwegen[2], and Dirk Schübeler[2]

1: equal contributions
2: corresponding authors

published in Genome Research, July 9, 2012

While changes in chromatin are integral to transcriptional reprogramming during cellular differentiation, it is currently unclear how chromatin modifications are targeted to specific loci. To systematically identify transcription factors (TFs) that can direct chromatin changes during cell fate decisions, we model the genome-wide dynamics of chromatin marks in terms of computationally predicted TF binding sites. By applying this computational approach to a time course of Polycomb-mediated H3K27me3 marks during neuronal differentiation of murine stem cells, we identify several motifs that likely regulate dynamics of this chromatin mark. Among these, the motifs bound by REST and by the SNAIL family of TFs are predicted to transiently recruit H3K27me3 in neuronal progenitors. We validate these predictions experimentally and show that absence of REST indeed causes loss of H3K27me3 at target promoters in trans, specifically at the neuronal progenitor state. Moreover, using targeted transgenic insertion, we show that promoter fragments containing REST or SNAIL binding sites are sufficient to recruit H3K27me3 in cis, while deletion of these sites results in loss of H3K27me3. These findings illustrate that the occurrence of TF binding sites can determine chromatin dynamics. Local determination of Polycomb activity by Rest and Snail motifs exemplifies such TF based regulation of chromatin. Furthermore, our results show that key TFs can be identified ab initio through computational modeling of epigenome datasets using a modeling approach that we make readily accessible.

2.1 Introduction

Cellular differentiation entails organized changes in gene expression. Pluripotent stem cells that commit to a somatic fate have to stably repress pluripotency genes and activate lineage specific genes in a temporally correct fashion. This regulation is coordinated by TFs in concert with dynamic changes in local chromatin organization of the DNA template. These changes have recently been documented in genome-wide analyses of histone modifications and DNA methylation (Mikkelsen et al. 2007; Meissner et al. 2008; Mohn et al. 2008; Zhou et al. 2011). Together with genetic studies epigenome maps have helped to establish the relevance of differentiation specific reprogramming of chromatin. While several large international efforts to gather epigenome data have been launched (Satterlee et al. 2010; Abbott 2011), only limited tools exist to determine the regulatory circuitry that guides chromatin dynamics.

Chromatin modifications can act upstream of TF activity by inhibiting or enhancing their ability to bind their cognate sites in the DNA (Barrera and Ren 2006; Kouzarides 2007). In turn, TFs can also act upstream of chromatin modifications by recruiting chromatin modifying enzymes, that modify the epigenome (Chan and La Thangue 2001; Lee et al. 2005). It is this latter mechanism that we wish to investigate here in a systematic manner. Since mammalian genomes encode an estimated 1500-2000 TFs (Vaquerizas et al. 2009), a comprehensive experimental investigation of all TFs is precluded, and other strategies are thus required to identify candidate TFs that are involved in particular aspects of chromatin regulation. To address this need, we adapted our recently published MARA (motif activity response analysis) approach, which models gene expression dynamics in terms of predicted transcription factor binding sites (TFBS) (Suzuki et al. 2009), to instead model genome-wide measured chromatin dynamics. The resulting Epi-MARA (Epigenome-motif activity response analysis) provides an analytical approach to identify TFs associated with chromatin reorganization ab initio, which we have made directly accessible through a web server (`http://ismara.unibas.ch`).

Here we use this approach to identify TFs that are involved in dynamic changes of a chromatin modification set by the Polycomb system, arguably the most relevant gene repression system during development (Schuettengruber and Cavalli 2009; Simon and Kingston 2009; Beisel and Paro 2011; Margueron and Reinberg 2011). A central component of Polycomb-mediated silencing is trimethylation of lysine 27 on histone H3 (H3K27me3), which is set by the Polycomb repressive complex 2 (PRC2). The protein enhancer of Zeste homologue 2 (EZH2) catalyses the methylation of H3K27 as part of PRC2 (Czermin et al. 2002; Muller et al. 2002) and it is required for differentiation and reprogramming (O'Carroll et al. 2001; Pereira et al. 2010). Target genes include important developmental regulators in mouse embryonic stem (ES) cells (Boyer et al. 2006) and are in part cell-type specific (Bracken et al. 2006; Mikkelsen et al. 2007; Mohn et al. 2008). Although DNA binding factors with limited sequence specificity have been implicated in targeting of the Polycomb system in flies (Ringrose and Paro 2007; Schwartz and Pirrotta 2008), the question of how Polycomb targets are specified remains currently unresolved, especially in vertebrates (Simon and Kingston 2009; Beisel and Paro 2011). We applied Epi-MARA on three developmental time points, where we measured the H3K27me3 chromatin mark during in vitro neurogenesis starting from murine stem cells (Mohn et al. 2008) resulting in a number of putative TFs involved in regulating different aspects of this chromatin reorganization. Among the top predictions were that binding sites for the SNAIL family of TFs and for sites of the TF RE-1 silencing transcription factor (REST) are associated with a transient increase in H3K27me3 at promoters in neuronal progenitors. We experimentally validate the role of REST using genome-wide mapping of REST binding and H3K27me3 levels in both wild type and REST-knockout cells. Furthermore, using transgenic constructs we show that promoter fragments containing REST or SNAIL binding sites are sufficient to recruit H3K27me3 while fragments in which these sites are deleted show reduced H3K27me3 levels. These results provide clear examples in vertebrates of TFs that locally influences Polycomb activity and, more generally, illustrate that TFs with regulatory function for epigenome reprogramming can be identified ab initio using computational modeling.

2.2 Results

2.2.1 Predicting mediators of chromatin changes using Epi-MARA

Comprehensive mapping of promoter regions (Harbers and Carninci 2005; de Hoon and Hayashizaki 2008; Balwierz et al. 2009) combined with comparative genomic prediction of TFBSs (van Nimwegen 2007) for known TF binding motifs (Wasserman and Sandelin 2004) have enabled researchers to study to what extent local occurrence of TFBSs can explain patterns of gene expression (Beer and Tavazoie 2004; Gao et al. 2004; Das et al. 2006; Suzuki et al. 2009). Here, we ask to what extent dynamic changes in chromatin can be explained by local TFBS occurrence and aim to identify the TFs that are involved in modulating chromatin locally.

To address this question systematically, we adapted our recently developed Motif Activity Response Analysis (MARA) (Suzuki et al. 2009), which models mRNA expression dynamics in terms of predicted TFBSs, to model genome-wide patterns of epigenetic marks, and termed this approach Epi-MARA (Fig. 1). Concretely, if M_{ps} quantifies the amount of a particular epigenetic mark M at promoter p in sample s, and N_{pm} denotes the total number of predicted binding sites for regulatory motif m in promoter p, then we assume a linear model of the following form:

$$M_{ps} = \text{noise} + \sum_m N_{pm} A_{ms}, \tag{2.1}$$

where c_p is the basal level of the chromatin mark at promoter p, and A_{ms} is the unknown activity of motif m in sample s, which is inferred by Epi-MARA (see Methods). Abstractly speaking, the activity A_{ms} quantifies how much each occurrence of motif m contributes to the level of epigenetic mark M in sample s. One can think of A_{ms} as reflecting the occupancy of TF binding to sites of motif m and the resulting effect on chromatin mark M. Thus, whenever Epi-MARA infers a highly positive activity A_{ms}, this predicts that the binding TF recruits the chromatin mark at stage s, whereas a highly negative A_{ms} implies that the binding TF inhibits deposition of the mark at stage s.

Notably, it is not the aim of Epi-MARA to provide accurate fits of epigenetic profiles at individual promoters. Since the actual levels of a chromatin mark at any promoter are likely a complex function of many variables acting both in cis and in trans the simple linear model of equation (1) typically captures only part of the variance in epigenetic mark levels. Importantly, however, the motif activities are inferred from the combined statistics of the hundreds to thousands of promoters that contain a given motif. Thus, the linear model applied by Epi-MARA effectively averages out the complications at individual promoters, and the remaining signal provides a robust statistical average activity for each motif, enabling reliable prediction of the TFs involved in chromatin mark dynamics. To allow easy application of this method, we have made automated Epi-MARA analysis available online (`http://ismara.unibas.ch`).

As a biological model of dynamic changes of transcriptome and epigenome we used a well-characterized mouse differentiation system, which progresses from embryonic stem (ES) cells to terminal neurons (TN) through a defined neuronal progenitor state (NP) (Bibel et al. 2004; Plachta et al. 2004; Bibel et al. 2007). We set out to identify the possible role of TFs in cell-type specific targeting of Polycomb-mediated H3K27 methylation in this system and applied Epi-MARA to our dataset of H3K27me3 at promoters in the ES, NP and TN stages (Mohn et al. 2008). The general approach is shown in Figure 1 together with the predicted activities of the nine motifs that contributed most to explaining the genome-wide H3K27me3 dynamics at promoters. Five of these nine, i.e. Sp1, Snail, Zeb1, Rest, and Arnt/Ahr, show a pattern in which there is a strong transient increase in motif activity at the NP stage. That is, Epi-MARA predicts the TFs binding these motifs to be involved in the recruitment of H3K27me3 going from the ES to NP stage. Of these candidate TFs we chose REST as a target for in-depth experimental validation as it is the only one of these motifs that is likely bound by a single TF and thus highly suitable for functional testing by genetic deletion. In contrast, Snail, Zeb1, and Sp1 motifs can each be recognized by multiple TFs (Postigo and Dean 2000; Bouwman and Philipsen 2002; Nieto 2002).

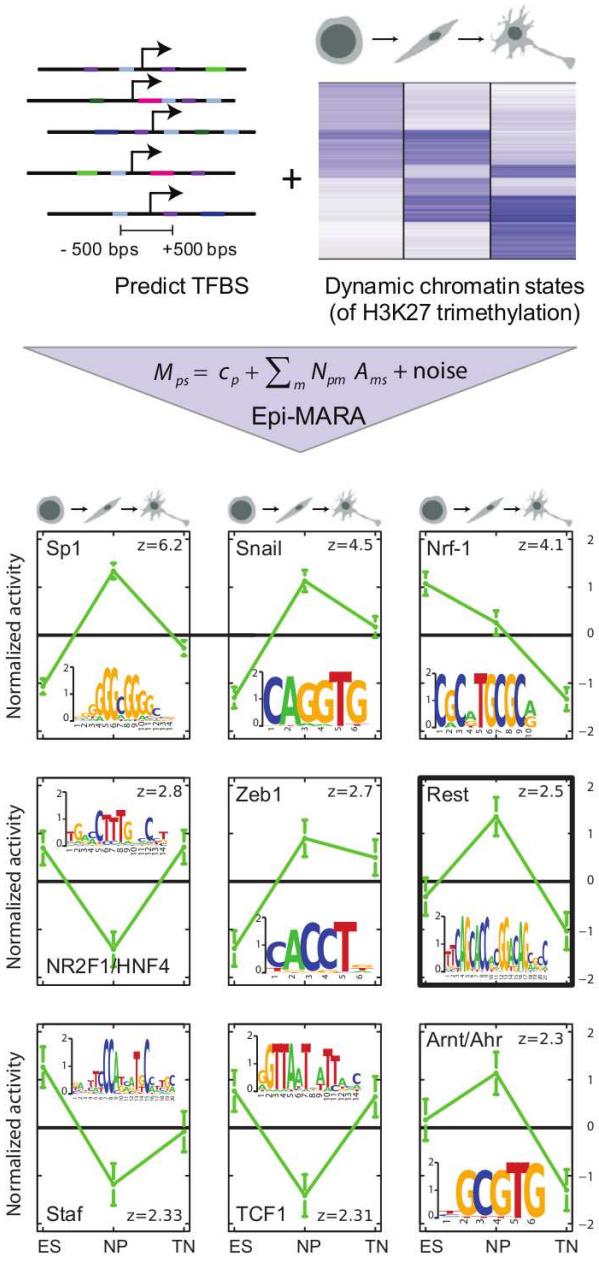


Figure 2.1: Epi-MARAs approach to predicting transcription factor activities that explain dynamics in H3K27me3 levels during neuronal differentiation: Transcription factor binding sites were predicted in proximal promoters genome-wide, using a Bayesian method that explicitly models binding site evolution. Epi-MARA models measured chromatin dynamics in terms of predicted TFBSs. M_{ps} quantifies the amount of a particular epigenetic mark M at promoter p in sample s, N_{pm} denotes the total number of predicted binding sites for regulatory motif m in promoter p, c_p indicates the basal level of the mark at promoter p and A_{ms} is the unknown activity of motif m in sample s. Depicted are the normalized activity profiles of the top nine motifs (green lines, with standard errors indicated) with their respective z-values. The three time points correspond to the embryonic stem cell (ES), neuronal progenitor (NP), and terminal neuron (TN) stage. Sequence logos of each of the motifs and the transcription factors thought to bind to them are shown as insets.

To compare the activity of TFs in regulating chromatin dynamics with their activities regulating expression we also analyzed transcriptome data of the three consecutive stages using the MARA method (Suzuki et al. 2009). One of the motifs that, according to the MARA analysis, most significantly regulates expression changes is the E2F motif (Supplementary Fig. 1). Its inferred transcriptional activity is highly positive in the ES and NP stages where cells are proliferating, while it strongly decreases at the TN stage where cells are post-mitotic and have exited the cell cycle. This is consistent with the known function of the E2F family of cell-cycle regulators that bind to this motif (Tao et al. 1997). In contrast Epi-MARA predicts no significant activity on H3K27me3 dynamics for the E2F motif.
Interestingly, the TF REST is also inferred to have an important role in driving expression changes, and its activity profile is consistent with its known role as a repressor of neuronal genes in non-neuronal tissues (Schoenherr and Anderson 1995). That is, REST target genes become active at the TN stage where REST itself is downregulated (Supplementary Fig. 2a). However, the activity profile of REST directing expression changes (Supplementary Fig. 1) is clearly distinct from its activity profile directing H3K27me3 (Fig. 1), suggesting that RESTs effects on transcription levels are at least partially independent from its effects on H3K27me3 levels. Notably, we find that predicted REST sites have higher than average H3K27me3 levels at all three time points in line with previous observation in individual cell states (Zheng et al. 2009; Liu et al. 2010) (Supplementary Table 1). We note that two recent studies, which appeared well past our decision to functionally validate the role of REST, reported biochemical interaction between REST and members of the Polycomb group (Ren and Kerppola 2011; Dietrich et al. 2012). However, these observations of a general co-occurrence of REST and Polycomb do not predict the cell-type specific activity for REST, which depends on the analysis of dynamic changes in H3K27me3 levels across the time course.

2.2.2 Experimentally determined REST binding sites support the computational prediction

To ask whether Epi-MARAs activity prediction, which is based on computationally predicted REST sites, is confirmed by REST binding sites that are indeed occupied by the factor, we mapped REST binding at the ES and NP stages. We carried out chromatin immunoprecipitation (ChIP) of REST bound DNA and subjected the precipitated DNA to high throughput sequencing (ChIP-seq). Peak finding was done on pooled replicates and revealed 1599 REST binding peaks in ES cells and 1035 in progenitors. Identified binding sites show a large overlap to those previously reported (Supplementary Fig. 3 and (Johnson et al. 2008)). The reduced number of peaks in progenitors likely reflects the fact that REST protein levels decrease during neuronal differentiation (Supplementary Fig. 2a). In agreement with this hypothesis 97% of the peaks present in progenitors are also present in stem cells. The majority of REST peaks contain a predicted binding site (Supplementary Table 2) and the number of predicted sites and the amount of binding as assayed by ChIP-seq correlate positively ($r = 0.48$, pvalue $2.9 \cdot 10^{53}$). Furthermore, REST binding occurs preferentially in proximity to transcription start sites (TSS) (Supplementary Fig. 2b) and we classified genes with REST binding within $\pm 2kb$ of the TSS as potentially regulated by this factor (Supplementary Fig. 2c). Mammalian promoters separate into two classes associated with either high or low density of CpG dinucleotides (Bird 1986; Carninci et al. 2006; Balwierz et al. 2009), and we observe that REST predominantly targets high-CpG promoters (Supplementary Table 3). Interestingly, promoter proximal REST binding sites show a distinct positioning immediately downstream of TSS (Sun et al. 2005; Zhang et al. 2006), which we also observe for both predicted and measured REST binding (Fig. 2a). While there is general agreement between predicted and measured REST binding, not all predicted promoter sites are occupied and some of the promoter proximal REST peaks lay just outside the regions covered by the computational predictions. We therefore asked whether Epi-MARA predicts different activities for REST if we replace the computationally predicted REST sites with the actual binding data (see Methods). This analysis resulted in a strikingly identical activity profile for REST, but with much larger significance as the z-value almost doubled (Fig. 2b). These results not only support the REST activity profile inferred using the TFBS predictions but also illustrate how actual in vivo binding data can be

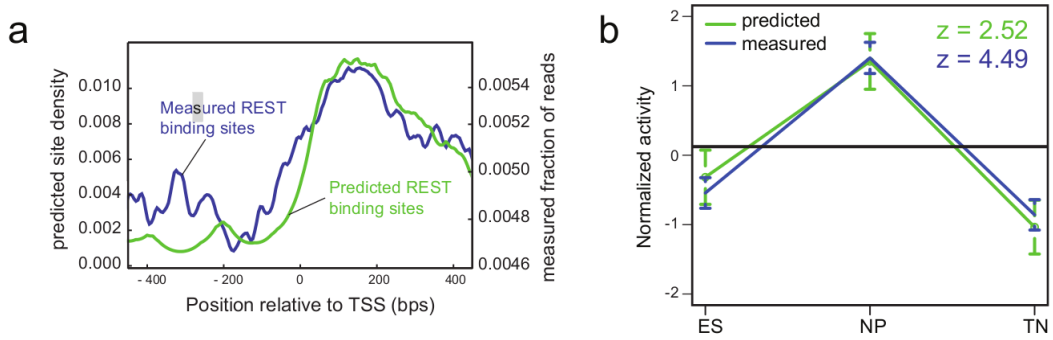


Figure 2.2: Analysis of REST binding data supports computational predictions: **a)** Frequency of predicted (green line) and measured (blue line) binding sites around transcription start sites. **b)** REST activity profiles calculated by Epi- MARA are similar when using either computationally predicted (green line) or measured REST binding sites (blue line). The prediction has higher significance when using the measured sites as indicated by the higher z-value (i.e. higher variance in activity relative to the error-bars).

incorporated, which, in this case, increased the accuracy of Epi-MARA's inference.

2.2.3 REST binding is associated with H3K27me3 dynamics genome-wide

Next, we assessed H3K27me3 dynamics beyond promoter regions by performing ChIP-seq at the three differentiation stages and determined all genomic regions that were enriched for H3K27me3 in at least one of the cellular states (see Methods). First, we noted that H3K27me3 levels peak immediately downstream of the TSS very similar to the binding pattern of REST (Supplementary Fig. 4a). Moreover, H3K27me3 levels peak around REST binding sites suggesting that the TF and chromatin mark co-localize (Supplementary Fig. 4b). If REST is mediating deposition of H3K27me3 by recruitment of Polycomb, we would expect REST binding to also co-localize with members of the PRC2 complex. To test this we analyzed SUZ12 ChIP-seq binding data from mouse embryonic stem cells (Pasini et al. 2010) and neuronal progenitors around REST binding sites. Importantly, we find that SUZ12 is even more localized at REST sites than H3K27me3 (Supplementary Fig. 4b).
Although many H3K27me3 enriched regions occur proximal to promoters, more than two thirds of H3K27me3 enriched regions are distal to promoters. However, these distal H3K27me3 regions are much less likely to be targeted by REST than promoter-proximal regions (Supplementary Table 4). Given RESTs preferred targeting to high-CpG promoters, we investigated the CpG content of all H3K27me3 regions and found that, strikingly, these separate into high- and low-CpG classes, similar to promoters (Fig. 3a). Moreover, CpG content cleanly distinguishes proximal and distal H3K27me3 regions, with 85% of proximal regions being high-CpG and 75% of distal regions being low-CpG (Fig. 3a). High-CpG regions are further distinct as they show higher levels of H3K27me3 than low-CpG regions (Supplementary Fig. 4c). Motivated by these differences, we asked whether Epi-MARA predicts different motif activities for REST if we analyze high- and low-CpG regions separately (see Methods). For high-CpG regions Epi-MARA predicts the same general activity profile for REST as previously for promoters, but with even higher significance (Fig. 3b). Strikingly, for low-CpG regions RESTs significance is not only reduced but the inferred activity is almost opposite to that of REST on high-CpG regions (Fig. 3b), i.e with a transient loss of H3K27me3 at the NP stage. Interestingly, high- and low-CpG regions have distinct H3K27me3 dynamics in general and the dynamics observed at REST targets are consistent with Epi-MARAs predictions (Fig. 3c, d).
In summary, genome-wide analysis of H3K27me3 levels predicts that REST binding at high-CpG regions,

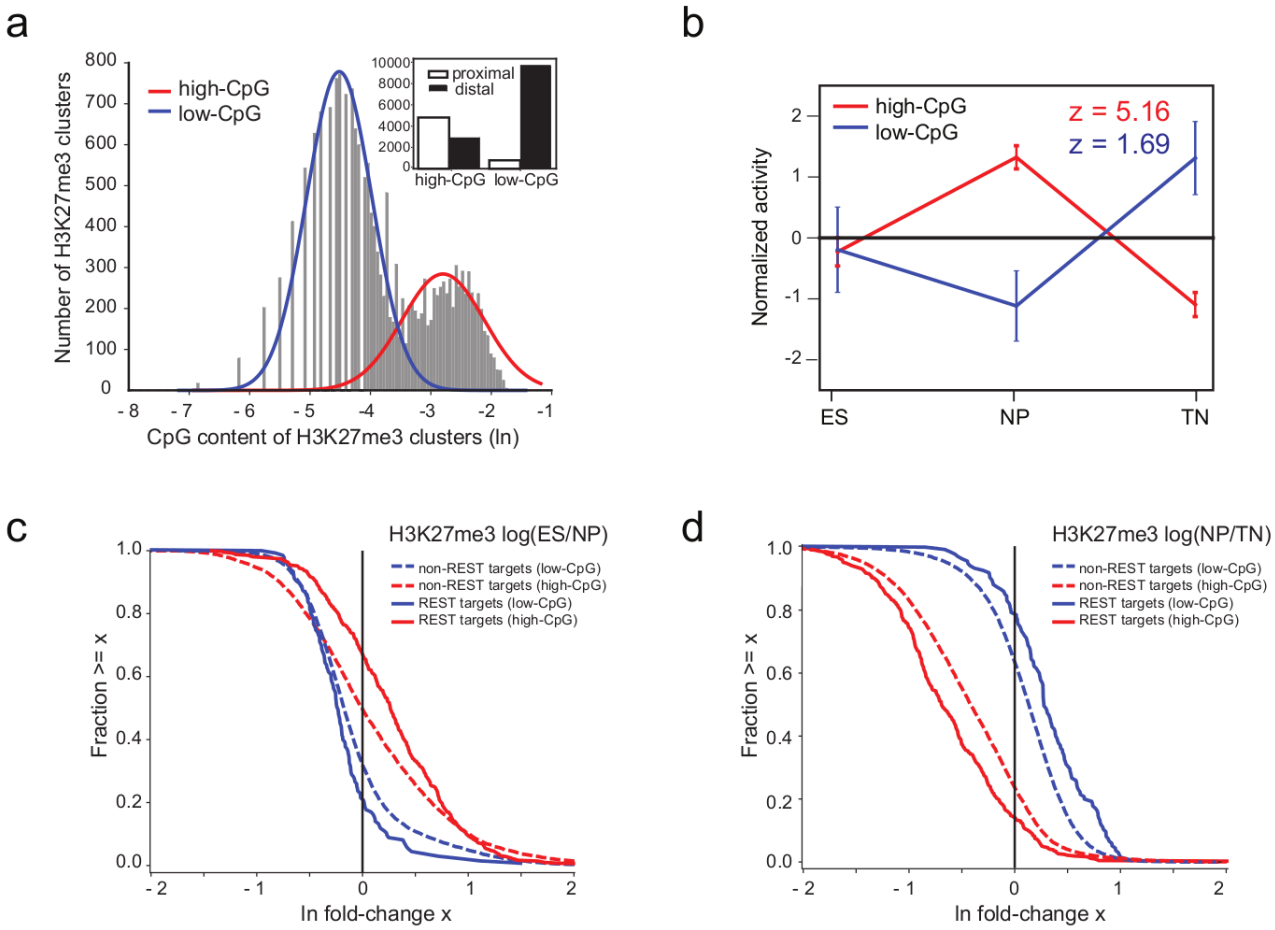


Figure 2.3: REST is associated with H3K27me3 dynamics at high- and low-CpG regions genome-wide: **a)** The distribution of CpG dinucleotide frequencies of H3K27me3 regions genome-wide is bimodal and can be fit by a mixture of two log-normal distributions (red and blue lines) corresponding to high- and low-CpG regions, respectively. Inset shows the numbers of K27me3 regions that are promoter-proximal and distal for high-CpG and low-CpG regions. **b)** REST activity profiles on high- (red) and low-CpG regions (blue) as inferred by running Epi-MARA on all H3K27me3 regions genome-wide show a transient gain and loss, respectively, at the NP stage. Note that, whereas REST activity on the high-CpG regions is highly significant, on the low-CpG regions REST activity has a much weaker significance. **c)** Reverse cumulative distributions of changes in H3K27me3 levels at the transition from ES to NP stage. We divided regions that were enriched for H3K27me3 into high-CpG/low-CpG (red/blue) and REST-target/non-target (solid/broken lines) regions. At high-CpG regions REST targets tend to gain H3K27me3 going from the ES to NP stage whereas non-target regions are equally likely to gain or lose H3K27me3. In contrast, most low-CpG regions lose H3K27me3 going to the NP stage and REST targets tend to lose even more H3K27me3. **d)** As in panel c but now for the transition from the NP to TN stage. High-CpG regions generally tend to lose H3K27me3 and REST targets tend to lose even more, whereas low-CpG regions tend to gain H3K27me3 and REST targets tend to gain even more.

which includes most promoter proximal REST targets, leads to a transient gain in H3K27me3 at the NP stage. In addition, a less significant transient loss of H3K27me3 at the NP stage for low-CpG regions is also predicted by Epi-MARA. We next tested these predictions by analyzing cells in which the Rest gene is deleted.

2.2.4 REST protein is required for local H3K27 methylation levels

REST is an essential protein for development as knockout mice die at embryonic day 11.5 (Chen et al. 1998). However, knockout ES cells (RESTko) are viable and show no defects in pluripotency (Jorgensen et al. 2009; Yamada et al. 2010) enabling us to test if they are competent to undergo neuronal differentiation in our in vitro system. Here, RESTko cells formed morphologically normal neurons with high efficiency, correct marker protein expression and limited changes in gene expression (Supplementary Fig. 5 and Supplementary Fig. 6), suggesting that REST is not essential for the initial steps of neuronal differentiation in vitro.
Next, we measured genome-wide H3K27me3 levels in RESTko cells at the stem cell and progenitor stages to investigate whether RESTs absence affects H3K27me3 levels at its target genes. We separated all regions enriched for H3K27me3 at any of the stages into high-CpG versus low-CpG and further into REST-target and non-target (see Methods). Next, we compared H3K27me3 levels in wildtype and RESTko cells between these four classes. This reveals little difference between REST target regions and non-target regions at the ES stage (Table 1 and Fig. 4b), in line with Epi-MARAs predicted REST activity at this stage. In contrast at the NP stage, as exemplified at two loci in Figure 4a, we observe a substantial loss of H3K27me3 in the RESTko cells relative to wildtype cells, affecting a substantial number of high-CpG REST targets (Table 1, Fig. 4b and Supplementary Fig.7). In addition, although the changes at low-CpG regions are much weaker, a notable gain of H3K27me3 is observed at low-CpG REST targets (Fig. 4b). This experimentally confirms Epi-MARAs predictions for REST at both high- and low-CpG regions. We conclude that REST contributes functionally to local levels of H3K27me3, which is strongest at high-CpG regions in neuronal progenitors. Next we tested if the observed loss of H3K27me3 is accompanied by a loss of PRC2, which mediates the H3K27me3 mark. We compared occupancy of the PRC2 component SUZ12 in RESTwt and RESTko neuronal progenitors. This reveals a loss of SUZ12 at a substantial number of high-CpG REST targets (Supplementary Fig.8a) and a loss of co-localization of SUZ12 with REST binding (Supplementary Fig. 8b). Moreover, compatible with a role for REST in Polycomb recruitment, there is a correlation between reduction in SUZ12 levels and reduction in K27me3 levels at high-CpG REST targets (Supplementary Fig.8c).

2.2.5 REST affects H3K27me3 and expression independently at many target genes

Since REST is an established repressor of gene activity it is conceivable that loss of H3K27me3 at proximal REST targets is a direct consequence of transcriptional upregulation. This would imply that all genes with REST-dependent loss of H3K27me3 are transcriptionally upregulated in RESTko cells. Although, as expected from a known repressive mark, there is a positive correlation between H3K27me3 loss and gene expression, this correlation is rather weak ($r = 0.28$ in ES and $r = 0.44$ in NP, Supplementary Fig. 9a). Most importantly, a third of the regions that lose H3K27me3 at the NP stage are not significantly transcriptionally upregulated (Supplementary Fig. 9). We thus conclude that the crosstalk between REST and the Polycomb pathway is independent of transcriptional changes at a substantial number of REST targets.

2.2.6 Promoter fragments containing REST or SNAIL binding sites locally recruit methylation of H3K27

Having established that absence of REST protein leads to a decrease of H3K27me3 at high-CpG binding sites, we wanted to further ask whether fragments of high-CpG promoter regions containing a REST site can recruit H3K27me3, and whether the REST binding site contributes to this recruitment. To this end we

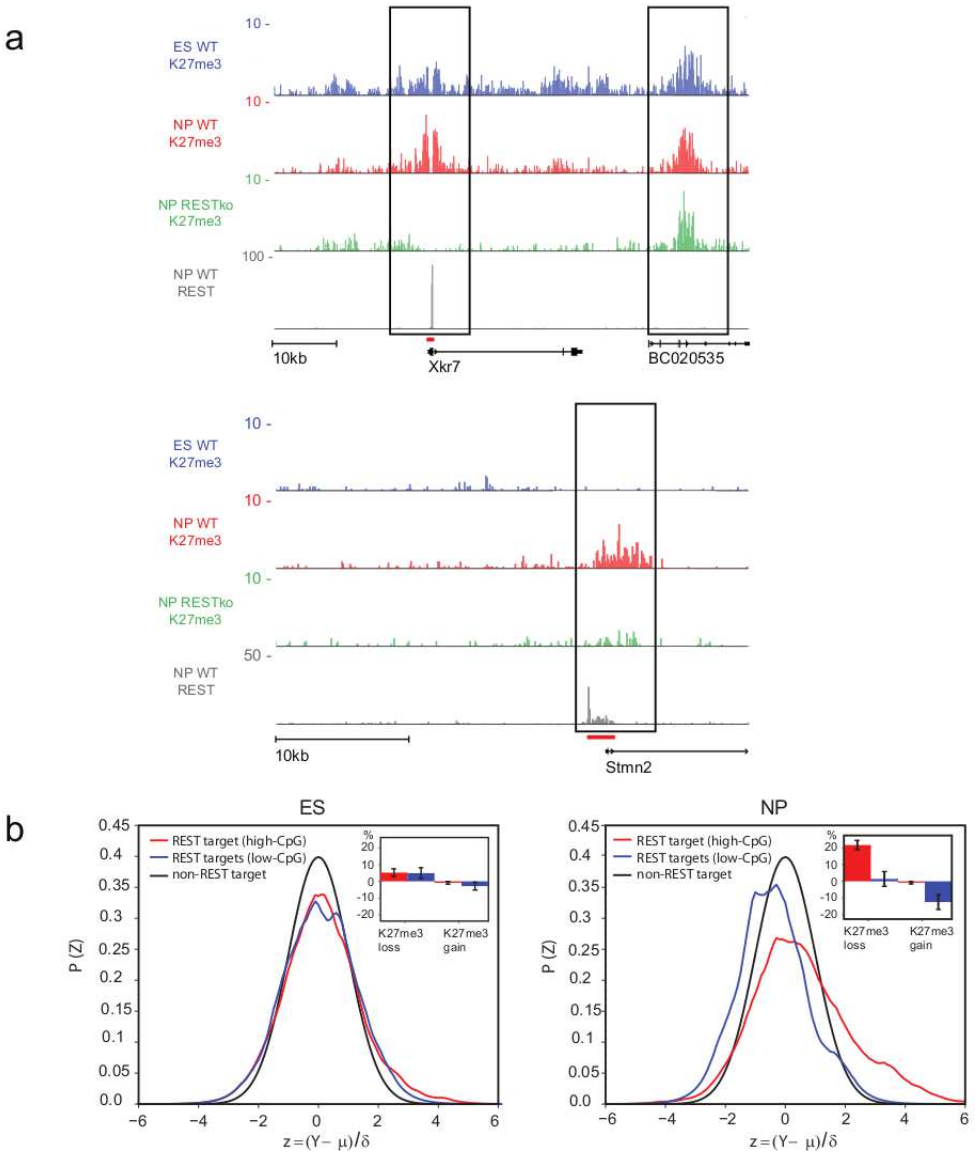


Figure 2.4: REST is required for H3K27me3 dynamics in neuronal progenitor cells: **a)** ChIP-Seq signal for H3K27me3 and REST in representative genomic regions. Shown are H3K27me3 signal in ES cells, NPs of wildtype (WT) and RESTko cells as well as REST signal in NPs. The top panel exemplifies selective loss of H3K27me3 at the REST binding site of the Xkr7 locus, whereas neighboring regions (BC020535) remain unaffected. The lower panel shows similar loss of H3K27me3 at the Stmn2 locus. Both the Xkr7 and Stmn2 locus are examples of promoter proximal high-CpG regions. Shown are normalized read densities. The red bars at the REST peaks indicate the regions cloned for transgenic experiments. **b)** Global comparison of H3K27me3 levels between WT and RESTko cells. Shown are the normalized distributions (Methods) of the ratio between H3K27me3 in WT versus RESTko for non-target regions (black lines) and for either low-CpG (blue lines) or high-CpG (red lines) regions that are REST targets at the ES (left panel) and NP (right panel) stage. The insets show the estimated fractions of REST targets that significantly lose or gain H3K27me3 in the RESTko at high-CpG (red) and low-CpG regions (blue). There are few significantly changing targets at the ES stage. At the NP stage a significant fraction of high-CpG targets lose H3K27me3 and a smaller but still significant fraction of low-CpG targets gain H3K27me3 in the RESTko cells.

generated reporter constructs consisting of 1.2 to 2kb promoter fragments containing a REST site, and mutant versions in which the REST site had been deleted. To ensure comparable chromatin organization we placed these sequence variants in wildtype cells into the same chromosomal locus using a Cre-recombinase based targeting system (Feng et al. 1999; Lienert et al. 2011). This site-specific targeting further enables us to control for genomic environment and thus to directly compare wildtype and mutant sequences (Fig. 5a). Importantly, the chosen "test site" is positioned within a genomic region that harbors no H3K27me3 and no REST binding (Lienert et al. 2011; Stadler et al. 2011). Thus, any REST or H3K27me3 signal should primarily reflect the recruitment abilities of the inserted sequence fragments. We inserted wildtype and mutated (REST) promoter fragments (Fig. 5b) of the following genes: Stmn2, Xkr7, Bdnf and Pgbd5. After targeted insertion and differentiation into neuronal progenitors we detect strong REST binding by ChIP to the wildtype, but no or weak binding in the four REST mutant sequences showing that the REST site is required for REST binding to the reporter constructs (Supplementary Fig. 10). Importantly, H3K27me3 is observed at all promoter fragments containing the REST site at the progenitor stage, whereas the mutant sequences show significant loss of H3K27me3 (Fig. 5c). In case of the Stathmin-like 2 (Stmn2) promoter, presence of the REST site results in a more than three-fold increase of H3K27me3 signal. Notably, the endogenous Stmn2 promoter shows no transcriptional response in RESTko cells. Of all four tested promoter fragments the Pgbd5 fragment shows the weakest loss of H3K27me3. Notably, the corresponding loss of REST binding at this promoter is also the weakest (Supplementary Fig. 10), suggesting that a cryptic binding site may still remain at this fragment. Together with the observed changes in H3K27me3 levels at genome-wide REST targets in the RESTko cells these results firmly establish that REST binding mediates Polycomb targeting and contributes to local levels of H3K27 methylation.

Besides REST, several factors that Epi-MARA predicted to play a role in H3K27me3 dynamics are recognized by a family of TFs. This makes loss of function approaches at the protein level very demanding. Our transgenic approach, however, can be used to assess the contribution of binding motifs to Polycomb recruitment irrespective of which TF from a family is binding. We thus extended our analysis to study the effect of the SNAIL binding site, another motif predicted to recruit K27me3 at the NP stage (Fig. 1). We inserted a total of six regulatory regions containing wildtype or mutated SNAIL sites (Fig. 5d) and tested for presence of H3K27 methylation. As seen with regulatory regions containing REST sites, we observe that all constructs containing SNAIL sites are sufficient to recruit H3K27me3. Deletion of the SNAIL sites leads to significant reduction of H3K27me3 for two of the three constructs tested (Fig. 5e). Notably, the construct that showed no significant response was the only one that contained only a single predicted SNAIL site, suggesting that the effect on H3K27me3 increases with the number of sites.

In summary, we conclude that promoter fragments containing bindings sites for SNAIL and REST TFs are sufficient to recruit H3K27me3 and, in line with the predictions, that these binding sites are a major contributor in cis to H3K27me3 levels.

2.3 Discussion

Recent genome-wide analyses of chromatin have established unexpected dynamics of the epigenome, which reflect cellular and developmental states. The analysis of such data has predominantly focused on characterizing the different kinds of chromatin domains that exist, and associating these domains with functional features such as active or inactive promoters or distal regulatory elements (Suzuki and Bird 2008; Meissner 2010; Ernst et al. 2011; Zhou et al. 2011). With the exception of chromatin modifications that are set by the process of transcription itself, such as H3K36 methylation, our understanding of how dynamic changes in chromatin are regulated remains limited. This likely reflects the complexity of the underlying targeting as different recruitment mechanisms for chromatin modifiers have been identified, including TFs, non-coding RNAs, as well as higher order nuclear organization (Schuettengruber and Cavalli 2009; Simon and Kingston 2009; Beisel and Paro 2011).

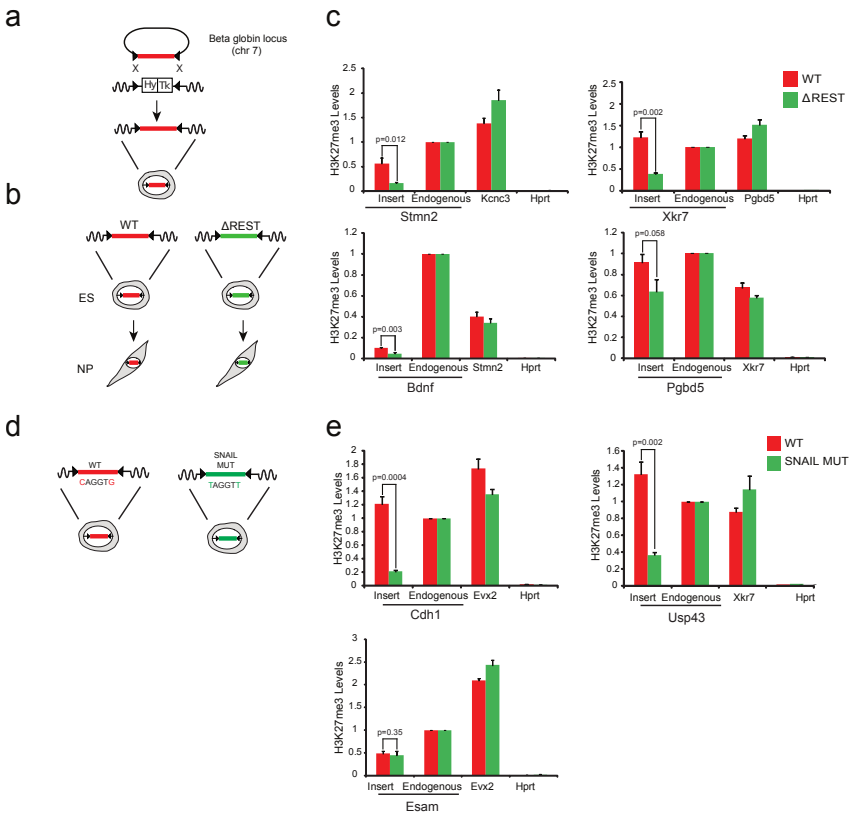


Figure 2.5: TFBS are required for H3K27me3 recruitment at the NP stage: a) Strategy to insert promoter regions into a defined genetic site (beta globin locus) via recombinase mediated cassette exchange (RMCE). The two marker genes inserted into the beta globin locus confer resistance against hygromycin (Hy) and sensitivity against ganciclovir (Tk), respectively and are flanked by two inverted lox sites (black triangles). Targeted insertion of a given transgene is achieved by Cre-mediated recombination and negative selection. b) The RMCE approach was used to insert several REST target promoter fragments with either wildtype sequence (WT) or REST site mutation (REST) into the beta globin locus. Correctly targeted ES cells were differentiated to the NP stage, where H3K27me3 and REST were measured at the inserted fragments. c) For each of the 4 inserts H3K27me3 levels were measured in cells bearing the WT fragment (red bars) and in cells bearing the REST fragment (green bars). Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. Note that different promoter regions are used as positive controls in the different panels. All inserted WT fragments show significant recruitment of H3K27me3 and loss in H3K27me3 for the ΔREST fragments. d) Either wildtype (WT) or mutated (MUT) promoter regions containing predicted SNAIL sites were inserted via RMCE. The SNAIL sites were mutated by changing the first and last nucleotide of the motif to a Thymidine. Correctly targeted ES cells were differentiated to the NP stage. e) For each of the 3 inserts H3K27me3 levels were measured in cells bearing the WT promoter (red bars) and in cells bearing promoters with mutated SNAIL sites (green bars). Note that the Cdh1, Usp43 and Esam promoter regions have three, two and one predicted/mutated SNAIL site, respectively. Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. All H3K27me3 levels are scaled to that of the endogenous region and error-bars show the standard error of three biological replicates. A p-value is shown and calculated for each insert using unpaired one-tailed t-test statistics.

2. MODELING OF EPIGENOME DYNAMICS IDENTIFIES TRANSCRIPTION FACTORS THAT MEDIATE POLYCOMB TARGETING

Here, we have tested the hypothesis that TFs contribute to dynamic changes in chromatin during cellular differentiation. We combined mapping of epigenetic marks at consecutive stages with computational modeling (Epi-MARA) to predict TFs involved in recruiting specific chromatin changes ab initio. We started from a data-set of murine embryonic stem cells undergoing neurogenesis, in which levels of H3K27me3 were measured at three consecutive cellular states during the differentiation. Application of Epi-MARA to this data identified several TFs as potential regulators of Polycomb dynamics during differentiation. Using several lines of evidence, we experimentally validate the prediction that REST is involved in transiently recruiting H3K27me3 to promoter regions at the neuronal progenitor stage: 1. Genetic deletion reveals that REST is necessary in trans for increased H3K27me3 levels at REST targets at the neuronal progenitor stage, specifically at high-CpG target regions, which includes almost all promoter proximal target regions. 2. Absence of REST causes loss of the PRC2 component SUZ12, mirroring the loss H3K27me3 at high-CpG regions. 3. Promoter fragments containing a REST binding site are sufficient in cis to recruit H3K27me3, whereas identical regions with mutated REST binding sites showed reduced recruitment. These findings support a model whereby local REST binding recruits Polycomb at the induction of in vitro neurogenesis.

Previous studies have already noted increased H3K27me3 signal at REST-bound promoters and enrichment of REST binding sites at CpG-islands bound by PRC2 (Ku et al. 2008; Zheng et al. 2009), while a more recent study showed that a non-coding RNA can bind to PRC2 and the LSD1/CoREST/REST complex in vitro (Tsai et al. 2010). During the preparation of this manuscript two studies reported biochemical interaction between REST and members of the PRC1 and PRC2 complexes (Ren and Kerppola 2011; Dietrich et al. 2012). Importantly however these correlative observations at single cell states did not identify the dynamic and context-dependent role of REST on H3K27 methylation that we predict based on chromatin dynamics and further validate experimentally. Notably, we show that absence of REST in stem cells has only subtle effects on H3K27me3 levels at target regions, suggesting that this previously noted co-occurrence of REST and H3K27me3 in stem cells has limited functional relevance. Importantly, and as predicted by our computational model, REST has more pronounced effects for H3K27me3 levels of target regions in neuronal progenitors (Fig. 4b).

While the detailed mechanisms of Polycomb targeting remain to be determined, our study suggests that, rather than a single dominant factor, it likely involves multiple TFs as we found several regulatory motifs associated with the increase of H3K27me3 at the neuronal progenitor stage. Of these, ZEB1 and the family of SNAIL factors bind to similar motifs and are important transcriptional repressors during epithelial-to-mesenchymal transition (Cano et al. 2000; Liu et al. 2008), which is compatible with a proposed function in Polycomb recruitment (Herranz et al. 2008). Here we experimentally confirm the role of SNAIL sites in regulating H3K27me3 levels using our transgenic approach (Fig. 5). Since Sp1 sites are among the most commonly occurring regulatory sites within CpG-islands, it is difficult to interpret whether the predicted role of Sp1 in H3K27me3 dynamics is specific to Sp1 or more generally associated with CpG-islands, which have been suggested to recruit PRC2 (Mendenhall et al. 2010). It is noteworthy, however, that Sp1-like sites are a component of Polycomb Responsive Elements (PRE) in Drosophila (Brown and Kassis 2010). In contrast, YY1, the mammalian ortholog of Pho, which is the most established TF with a function in Polycomb recruitment in Drosophila melanogaster, is unlikely to have that role in mammals (Ku et al. 2008; Mendenhall et al. 2010), at least in stem cells.

Based on recent work in Drosophila (Enderle et al. 2010) and mouse stem cells (Landeira et al. 2010; Brookes et al. 2012), it has been suggested that Polycomb might repress by stalling polymerases. Our observation that the Rest, Snail, and Zeb1 motifs tend to be positioned immediately downstream of TSS (Supplementary Fig. 11) is compatible with this model. However, this observation does not generally apply to the top 9 predicted motifs (Supplementary Fig. 11). We further show that the dynamics of H3K27me3 are different for high-CpG and low-CpG regions in line with a proposed model that local CpG richness influences Polycomb recruitment (Mendenhall et al. 2010; Lynch et al. 2011). We propose that this is connected to individual TF activity since REST has the strongest effect on H3K27me3 levels at high-CpG regions and a weaker opposite effect at low-CpG regions (Fig. 4b). This opposite behavior at high-CpG and low-CpG motifs does not gen-

eralize to all TFs (Supplementary Table 5).
Our results are compatible with a role for cell-type specific co-factors since the effect of REST on H3K27me3 are by far strongest at the NP stage, whereas REST binding decreases from the ES to NP stage. Nevertheless, the needed regulatory information can be highly localized as tested promoter fragments were sufficient to recruit H3K27me3 when inserted into a defined genomic region. While it remains to be seen if these elements fulfill the definition of a PRE, i.e. whether they repress genes in cis in a Polycomb-dependent fashion, our results suggest that both REST and SNAIL sites could contribute to such function. Clearly, Rest and Snail provide convincing examples for DNA binding motifs that enhance local Polycomb states in the mammalian genome. Epi-MARA provides a general methodology for inferring the stage-specific activities of TFs associated with chromatin dynamics that we foresee will be useful for the study of epigenome maps particularly in light of the multitude of datasets that are being generated as part of large epigenome initiatives (Satterlee et al. 2010; Abbott 2011). The approach makes use of sophisticated comparative genomic TFBS predictions and linear modeling, which accounts for the contributions of all regulatory motifs at once. The ability to predict TFs involved in regulating chromatin dynamics from epigenome data-sets provides a powerful tool in this context, as predicted TFs can be immediately subjected to follow-up experiments. The identification of the context-dependent function of REST, and the role of the SNAIL sites, illustrate its utility. Epi-MARA is directly accessible by our web server implementation (`http://www.mara.unibas.ch/cgi/mara_dev`).
Importantly, our findings have direct implications for regulatory models of chromatin regulation. In our neurogenesis system, a linear model in terms of predicted binding sites explains roughly the same fraction of variance in H3K27me3 at promoters as it explains variance in transcript levels (Supplementary Table 6). This result suggests that, like regulation of transcription, chromatin dynamics of H3K27me3 are regulated to a significant extent by local DNA sequence motifs that are recognized by trans-acting factors.

2.4 Acknowledgements

We thank Robert Ivanek, Lukas Burger, and Nacho Molina for advice, Nicolas Thoma, Susan Gasser and members of the Schbeler and van Nimwegen labs for comments on the manuscript and the laboratory for Quantitative Genomics of the ETH Zrich in Basel for next generation sequencing. We also thank David Anderson for providing one of the used REST antibodies.
Research in the laboratory of DS is supported by the Novartis Research Foundation, by the European Union (NoE EpiGeneSys FP7-HEALTH- 2010-257082, LSHG-CT-2006-037415), the European Research Council (ERC-204264), the Swiss National Science Foundation (Sinergia program) and the EMBO Young Investigator program. EvN acknowledges support by the Swiss National Science Foundation and the Swiss Institute of Bioinformatics. DS and EvN are both supported by the Swiss Systems Biology Initiative SystemsX.ch within the network ”Cellplasticity”.

Class and Stage	Percentage that lose K27me3	Percentage that gain K27me3
ES low-CpG	$4.9\% \pm 3.2\%$	$2.9\% \pm 2.3\%$
ES high-CpG	$5.2\% \pm 2.3\%$	$0.9\% \pm 0.8\%$
NP low-CpG	$1.4\% \pm 4.4\%$	$12.4\% \pm 4.4\%$
NP high-CpG	$21.7\% \pm 2.8\%$	$0.8\% \pm 0.7\%$

Table 1: Estimated percentages of REST targets that significantly lose/gain H3K27me3 in the RESTko cells, separately at low- and high-CpG regions, and separately at the ES and NP stages. Using as a cut-off targets that change by more than one standard-deviation (z¿1 and z¡-1 for loss and gain, respectively) we conservatively estimated the fraction of true targets as the percentage of REST targets with a z-value larger than the cut-off in addition to the percentage expected by chance based on the standard-normal distribution. Error bars are based on a Bayesian inference procedure (see Methods). Note that the largest percentage of true targets is observed for high-CpG regions losing H3K27e3 at the NP stage, followed by low-CpG regions gaining H3K27me3 at the same stage.

2.5 Methods

2.5.1 Epi-MARA

We here describe the main methods employed in the Epi-MARA analysis. Further details are supplied in the Supplementary methods. Epi-MARA models the dynamics of epigenetic marks in terms of predicted TFBSs in regulatory regions genome-wide, building on the Motif Activity Response Analysis that we developed previously[30]. Briefly, for each promoter we constructed multiple alignments using orthologous sequences from mouse, human, rhesus macaque, dog, cow, horse, and opossum, of the proximal promoter region consisting of 500 base pairs both upstream and downstream of the cluster of transcription start sites that defines the promoter[66]. Using databases of experimentally determined binding sites[67,68], we collected a set of 207 mammalian regulatory motifs (position specific weight matrices) representing the binding specificities of approximately 350 mammalian TFs. Then, using a Bayesian probabilistic method that explicitly models the evolution of TFBSs, we predict binding sites for all regulatory motifs in all proximal promoter regions[26]. We summarize the binding site predictions by a matrix with components N_{pm}, denoting the sum of the posterior probabilities of all binding sites for motif m in promoter p, which we also refer to as the 'number' of binding sites for motif m in promoter p. For each promoter and at each time point we quantify the occurrence of an epigenetic mark of interest by either the log-intensity of probes that lie within the promoter (when using ChIP-chip data) or the log-fraction of all sequence reads in a $4kb$ region centered on the promoter(when using ChIP-seq data). We denote this occurrence of the epigenetic mark by M_{pt} and assume the following linear model:

$$M_{pt} = \text{noise} + \sum_m N_{pm} A_{mt}, \tag{2.2}$$

where c_p is the basal level of the chromatin mark, and A_{mt} is the unknown activity of motif m at time point t. Using a Bayesian probabilistic frame-work, we then calculate a joint posterior probability distribution for all motif activities. To this end, we assume that the deviation between model and measured level M_{pt} (i.e. the 'noise' term in the above formula) is Gaussian distributed at each promoter and at each time point. In addition, to avoid over-fitting, we use a Gaussian prior on the activities A_{mt}, and we determine the variance of this prior by a cross-validation procedure. Finally, we infer both the maximal posterior activities and their standard-errors. To rank motifs, we measure the importance of a motif in explaining expression variations by a score similar to a z-statistic. The z-score z_m of motif m is quantified as an average squared z-value of the activity across conditions, i.e.

$$z_m = \sqrt{\frac{\sum_t \left(\frac{A^*_{mt}}{\sigma_{mt}}\right)^2}{T}}, \tag{2.3}$$

where T is the number of time points. Note that our z-scores are meant to rank the importance of motifs and cannot be used to assess the statistical significance of motif activities. To assess statistical significance of the motif activities that we observed, we performed the following permutation test: We randomly permuted the association between binding site predictions and promoters and reran Epi-MARA 1000 times, reporting the z-scores of the inferred motif activities for all 207 motifs in each run. Since in the 1000 permutations there was only one motif in one run with a z-score larger than 2.52, we infer that the probability of obtaining a z-score as high as $z = 2.52$ (the z-score of the REST motif on the ChIP-chip data) is approximately $p = 5 \cdot 10^{-6}$. To run Epi-MARA on all H3K27me3 enriched regions genome-wide we predicted TFBSs across the entire 4kb sequence of each H3K27me3 region using the same procedure as used for predicting sites in proximal promoters. For each H3K27me3 region we then determined which 1kb window contains the highest number of predicted binding sites (pooling all motifs) and we used the predicted sites within this 1kb region for the entries in the site-count matrix N_{pm} for the corresponding H3K27me3 region. To infer motif activities separately for high- and low-CpG regions we treat, for each motif m, sites within low-CpG

regions and sites within high-CpG regions as if they derived from two separate motifs, effectively doubling the number of motifs for which we infer activities.

2.5.2 Cell Culture

Wildtype mouse embryonic stem cells were derived from blastocysts (3.5 PC) of mixed 129-C57Bl/6 background and cultivated on feeder cells ($37C$, 7% CO2). REST knock-out and corresponding wildtype cells were obtained from Helle Jorgensen[48,49]. Differentiation of cells was performed as described previously[1,31].

2.5.3 Western Blot Analysis

For detection of REST protein levels during differentiation the total cell lysates of wildtype and REST knock-out cells were used for western blot analysis. The membrane was probed with mouse anti-REST (12C11, gift from David Anderson) and rat anti-tubulin (tissue culture supernatant, cell line YL1/2, ECACC) in combination with appropriate secondary antibodies coupled to HRP.

2.5.4 Immunocytochemistry

Cells were fixed with 2% paraformaldehyde, either three hours or 10 days after plating, and probed with mouse anti-Pax6 (chick PAX6 a.a 1-223, DSHB), rabbit anti-Nestin (Sigma N5413) and mouse anti-Tuj1 (MMS-435P, Covance). Proteins were detected by an appropriate secondary antibody conjugated to Alexa Flour.

2.5.5 Chromatin-IP

Cells were cross-linked in medium containing 1% formaldehyde for 10 min at room temperature. ChIP was carried out as previously described 69,70 with slight modifications. Antibodies used were -H3K27me3 (Millipore,#07-449) and -REST (Santa Cruz, #H-290). Chromatin was sonicated for 10 (stem cells) or 18 cycles (neuronal progenitors) of 30 sec using a Diagenode Bioruptor. Precipitated DNA was either analyzed by quantitative real time PCR or subjected to next generation sequencing.

2.5.6 Quantitative real time PCR

Real time PCR was performed using SYBR green chemistry (ABI). 1/40 of ChIP sample or 40 ng of input chromatin were used per PCR reaction. Primer sequences are available upon request. All data is shown with standard error from three biological replicates. Significances were calculated using unpaired 1-tailed students t-test statistics.

2.5.7 Next generation sequencing

5 to 10 ng of precipitated DNA was prepared for Solexa Sequencing as described 2. Briefly, ChIP DNA was ligated to adapters and ligation products of about 250 bp were gel purified on 1.5% agarose to remove unligated adaptors. DNA was amplified by 18 PCR cycles. DNA sequencing was carried out using the Illumina/Solexa Genome Analyzer II (GA2) sequencing system. All generated datasets are available for download at the GEO database using the following URLs: `www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=jxchzqgousyyeny&acc=GSE27148` and `www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=vzqbzucsasqwgpq&acc=GSE27114`.

2.5.8 Genomic coordinates

The July 2007 M. musculus genome assembly (NCBI37/mm9) provided by NCBI (`http://www.ncbi.nlm.nih.gov/genome/guide/mouse/`) and the Mouse Genome Sequencing Consortium (`http://www.sanger.ac.uk/Projects/M_musculus/`) was used as a basis for all analyses. Annotation of known RefSeq transcripts was obtained from UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/refGene.txt.gz`).

2.5.9 Read filtering, alignment and weighting

Low-complexity reads were filtered out based on their dinucleotide entropy as follows: For each read, the dinucleotide entropy was calculated according to the formula $H = i \cdot f_i \cdot \log f_i$, where f_i is the frequency of dinucleotide i in the read and the sum is over all dinucleotides (i from 1 to 16). The read was filter out if its H was less than half the dinucleotide entropy of the genome, typically removing less than 0.5% of the reads in a given sample. Alignments to the mouse genome were performed by the software bowtie (version 0.9.9.1)[71] with parameters -v 2 -a -m 100, tracking up to 100 best alignment positions per query and allowing at most two mismatches. To track genomically untemplated hits (e.g., exon-exon junctions or missing parts in the current assembly), the reads were also mapped to an annotation database containing known mouse sequences (miRNA from `ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/13.0`, rRNA,snRNA, snoRNA and RefSeq mRNA from GenBank `http://www.ncbi.nlm.nih.gov/sites/entrez`, downloaded on July 16, 2009, tRNA from `http://lowelab.ucsc.edu/GtRNAdb/`and piRNA from NCBI (accessions DQ539889 to DQ569912). In that case, all best hits with at most two mismatches were tracked. Each alignment was weighted by the inverse of the number of hits. In the cases where a read had more hits to an individual sequence from the annotation database than to the whole genome, the former number of hits was selected to ensure that the total weight of a read does not exceed one. All quantifications were based on weighted alignments. For generation of wiggle files samples were normalized for library size first and files were generated with a window size of 100bps.

2.5.10 Analysis of sequencing data

In order to detect REST peaks from the ChIP-seq data we slide a window of length 1000 bps along the genome and calculate, for each window, the fraction of all ChIP-seq reads from the REST IP and the fraction of all reads from a background sample (input DNA) that map to the window (since background counts are generally smaller, we use a window of 2000bps centered at the same position to obtain more robust background frequencies). Inspecting the reverse-cumulative distribution of background counts across the genome, we observed that a small subset of windows showed aberrantly high background frequencies (Supplementary Fig. 2.10b) and these windows were removed from further consideration (these windows typically correspond to regions with repeats that presumably occur more frequently in the genome of the cells from which our DNA was taken, than in the mm9 genome assembly). We assume that the noise in the estimated and follow Poisson distributions and calculate, for each window, a z-statistic:

$$z = \frac{f_{IP} - f_{bg}}{\sqrt{\frac{f_{IP}}{N_{IP}} + \frac{f_{bg}}{N_{bg}}}}, \tag{2.4}$$

where N_{IP} and N_{bg} are the total numbers of reads in the IP and background sample, respectively. Inspecting the reverse-cumulative distribution of z-statistics across the genome, we observe a long tail of highly enriched regions to the right of $z = 3.1$ (Supplementary Fig. 2.10a) and we denote all regions with consecutive windows with z-values larger than 3.1 as REST binding regions. To determine the false discovery rate of binding region prediction at this cut-off we made use of the fact that we measured the background distribution in duplicate and performed binding region prediction in the exact same way, treating one of the background

samples as a 'foreground' sample. From this we find that, by chance, a fraction $1.2 \cdot 10^{-4}$ windows genome-wide have a z-value larger than 3.1, leading to a total of 289 falsely discovered binding regions genome-wide, which should be compared to the 1624 REST binding regions determined from the REST IP sample. Any REST binding region whose center is less than 2kb from a known transcription start site (from the RefSeq collection of transcripts) is considered proximal. All other REST binding regions are considered distal.

To predict REST binding sites for all REST binding regions we again produced multiple alignments of orthologous regions from mouse, human, rhesus macaque, dog, cow, horse, and opossum, and ran the MotEvo algorithm 26 on each multiple alignment. We also searched for non-canonical sites of arbitrary spacing between the two half-sites of the REST motif. In contrast to previous work 42 which found only non-canonical sites with a spacer of 6bps, we find non-canonical sites with spacers of both 6 and 7 bps. Linear regression between the total number of predicted REST binding sites (i.e. the sum of posteriors of all predicted sites) at each REST binding region, and the binding z-statistic of the region, shows a correlation of $r = 0.48$ (p-value $2.9 \cdot 10^{-53}$). We compared our predicted REST binding regions with those of Johnson et al.[41] by collecting all regions they report and, for each region, calculating a REST fold-change enrichment of ChIP-seq reads. We then calculated the overlap of the predicted binding regions for a fold-change enrichment of 2 (Supplementary Fig. 2.3).

To obtain positional profiles with respect to TSS for the predicted REST binding sites, we summed the posteriors of all predicted binding sites at promoters at each position relative to TSS. To obtain positional profiles for the REST binding data and H3K27me3 signals we simply summed all reads from the corresponding IP samples at each position relative to TSS.

To perform Epi-MARA analysis with the REST binding data replacing REST binding site predictions we replace the predicted binding site counts with results of the REST binding assay at each promoter p. Since the z-statistics of REST binding at promoters have a very different distribution of values from those of the site counts N_{pm}, it is necessary to normalize the matrix N_{pm} such that binding site predictions and binding data can be quantitatively compared. We therefore replace the matrix N_{pm} with a binary matrix B_{pm} in which $B_{pm} = 1$ whenever $N_{pm} > 0.2$ and $B_{pm} = 0$ otherwise. Finally, we replace the column with one based on the REST binding data, i.e. where whenever there was a REST binding peak within 2kb of the corresponding promoter. For the analysis of the H3K27me3 data we slide a 2000bp window along the genome and calculate a z-statistic for each window quantifying the enrichment of the signal in the IP over the background sample by a z-statistic as above, pooling the data from the replicates and from the different stages. As for the REST binding data, we observe a long tail of high z-values which occurs for the H3K27me3 data to the right of $z = 4.0$ (Supplementary Fig. 2.10c) and we defined H3K27me3 enriched regions as consecutive windows that all have a z-statistic larger than 4.0. Using again the two background samples (Supplementary Fig. 2.10d) to determine a false discovery rate we find that, by chance, a fraction of $2.8 \cdot 10^{-4}$ of windows have a z-value larger than 4.0, leading to 427 false predicted H3K27me3 regions genome-wide, which should be compared to the 18293 regions predicted from the H3K27me3 IP samples. The H3K27me3 enriched regions are divided into different classes using a number of criteria. Regions that overlapped a proximal promoter, i.e. a promoter plus and minus 500bps, were considered proximal and all others were considered distal. Similarly, regions that overlap a REST binding peak were considered REST targets and all others non-targets. For each region enriched in H3K27me3 we slide a 1kb window over the region and calculated the CpG dinucleotide frequency within each window. We defined the CpG-content of a region as the highest CpG frequency of a 1kb window within it. Inspection of the distribution of log-CpG content across H3K27me3 enriched regions show two classes and we fitted the distribution of log-CpG content by a mixture of two Gaussians (Fig. 2.3a). After fitting of the Gaussian mixture, posterior probabilities for each region to belong to the high-CpG or low-CpG class were calculated in the standard Bayesian way. In subsequent analyses, distributions for low-CpG and high-CpG regions were obtained by weighing each region with the posterior probability that it belongs to the

corresponding class.

For each region that was enriched for H3K27me3 at any of the stages, we calculated log-fold changes between ES and NP and between NP and TN stages by calculating the log-ratios of the fractions of reads from the corresponding IP samples mapping to each of the regions.

To compare H3K27me3 levels between wildtype (WT) and RESTko mutant (KO) cells we collected all regions that were enriched for H3K27me3 in the wildtype cells at any of the stages. For each region we calculated the fractions f_{WT} and f_{KO} of all IP reads that mapped to that region in WT and KO and calculated both the absolute intensity $X = (\log f_{WT} + \log f_{KO})$ (summed over all replicates) as well as the log-ratio: $Y = \log \frac{f_{WT}}{f_{KO}}$ (averaged over the replicates). Supplementary Figure 2.9b shows, as a function of absolute intensity X, the average and standard error of Y for all regions that are non REST targets (black dots with error-bars) for both high-CpG and low-CpG regions at both the ES and NP stages. As these figures make clear, there are some systematic differences in the overall distribution of H3K27me3 signals between wildtype and the RESTko cells. Therefore, in order to properly compare H3K27me3 signals between wildtype and RESTko, we adopted a normalization procedure similar to Loess normalization. For each stage, we sorted all non-target regions by their absolute intensity X (averaging wildtype and RESTko intensities). For each region we then collected the 50 regions with values of X immediately below, and the 50 regions with values of X immediately above, and calculated the mean μ and standard deviation σ. In this way we estimated the expected mean μ and standard-deviation σ of non-targets, as a function of their absolute H3K27me3 levels. For each REST target we determined both its fold-change Y and absolute H3K27me3 level X and calculated a z-value $z = \frac{Y-\mu}{\sigma}$ using the expected mean and standard deviation of non-targets with absolute levels of H3K27me3 of X. To suppress fluctuations we averaged the z-statistics with a Gaussian kernel. Note that, per definition, the z-values of non-target regions follow a Gaussian distribution of mean zero and standard-deviation one. To estimate the fraction ρ of REST targets that significantly change H3K27me3 we compared the fraction of REST targets that show z-values more than one standard deviation away from the mean (i.e. $z > 1$ when considering targets losing H3K27me3 and $z < -1$ when considering targets gaining H3K27me3) with the fraction expected by chance using a Bayesian procedure. Let q denote the probability to obtain a z-value larger than 1 by chance according to the standard Gaussian. Conservatively assuming that all true targets must have a z-value larger than 1, the probability for a randomly chosen target to have a z-value larger than one is $p = \rho + (1 - \rho)q$. Given that there are N REST targets in total, of which n have a z-value larger than 1 we use Bayes' theorem to calculate a posterior probability distribution over ρ and estimate its mean and standard-deviation. We similarly estimate the fraction of targets that significantly gain H3K27me3, separately for each stage, and separately for t high- and low-CpG target regions.

2.5.11 RNA preparation and expression analysis

Total RNA was prepared using TRIzol (Invitrogen). mRNA expression data were generated using Mouse Gene 1.0 ST and Mouse Genome 430 2.0 arrays. Microarrays were RMA-normalized using R/Bioconductor[72] and the oligo package version 1.14.0[73]. To determine transcriptional regulation of REST target genes in the RESTko we selected a 2-fold change as cut-off for significant upregulation.

2.5.12 Recombinase mediated cassette exchange (RMCE)

2kb promoter fragments of REST targets were cloned and stably integrated into stem cells via RMCE as described (Lienert et al, submitted)[54]. ΔRE-1 binding site mutants were generated by removing 15 to 20bps of the RE-1 consensus sequence. Primer sequences are available upon request.

2.6 Bibliography

1. Abbott, A. 2011. Europe to map the human epigenome. Nature 477(7366): 518.
2. Balwierz, P.J., Carninci, P., Daub, C.O., Kawai, J., Hayashizaki, Y., Van Belle, W., Beisel, C., and van Nimwegen, E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol 10(7): R79.
3. Barrera, L.O. and Ren, B. 2006. The transcriptional regulatory code of eukaryotic cells–insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr Opin Cell Biol 18(3): 291-298.
4. Beer, M.A. and Tavazoie, S. 2004. Predicting gene expression from sequence. Cell 117(2): 185-198.
5. Beisel, C. and Paro, R. 2011. Silencing chromatin: comparing modes and mechanisms. Nat Rev Genet 12(2): 123-135.
6. Bibel, M., Richter, J., Lacroix, E., and Barde, Y.A. 2007. Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. Nat Protoc 2(5): 1034-1043.
7. Bibel, M., Richter, J., Schrenk, K., Tucker, K.L., Staiger, V., Korte, M., Goetz, M., and Barde, Y.A. 2004. Differentiation of mouse embryonic stem cells into a defined neuronal lineage. Nat Neurosci 7(9): 1003-1009.
8. Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. Nature 321(6067): 209-213.
9. Bouwman, P. and Philipsen, S. 2002. Regulation of the activity of Sp1-related transcription factors. Mol Cell Endocrinol 195(1-2): 27-38.
10. Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K. et al. 2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441(7091): 349-353.
11. Bracken, A.P., Dietrich, N., Pasini, D., Hansen, K.H., and Helin, K. 2006. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev 20(9): 1123-1136.
12. Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., Kimura, H. et al. 2012. Polycomb Associates Genome-wide with a Specific RNA Polymerase II Variant, and Regulates Metabolic Genes in ESCs. Cell Stem Cell 10(2): 157-170.
13. Brown, J.L. and Kassis, J.A. 2010. Spps, a Drosophila Sp1/KLF family member, binds to PREs and is required for PRE activity late in development. Development 137(15): 2597-2602.
14. Cano, A., Perez-Moreno, M.A., Rodrigo, I., Locascio, A., Blanco, M.J., del Barrio, M.G., Portillo, F., and Nieto, M.A. 2000. The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nat Cell Biol 2(2): 76-83.
15. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C. et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38(6): 626-635.
16. Carvalho, B.S. and Irizarry, R.A. 2010. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26(19): 2363-2367.

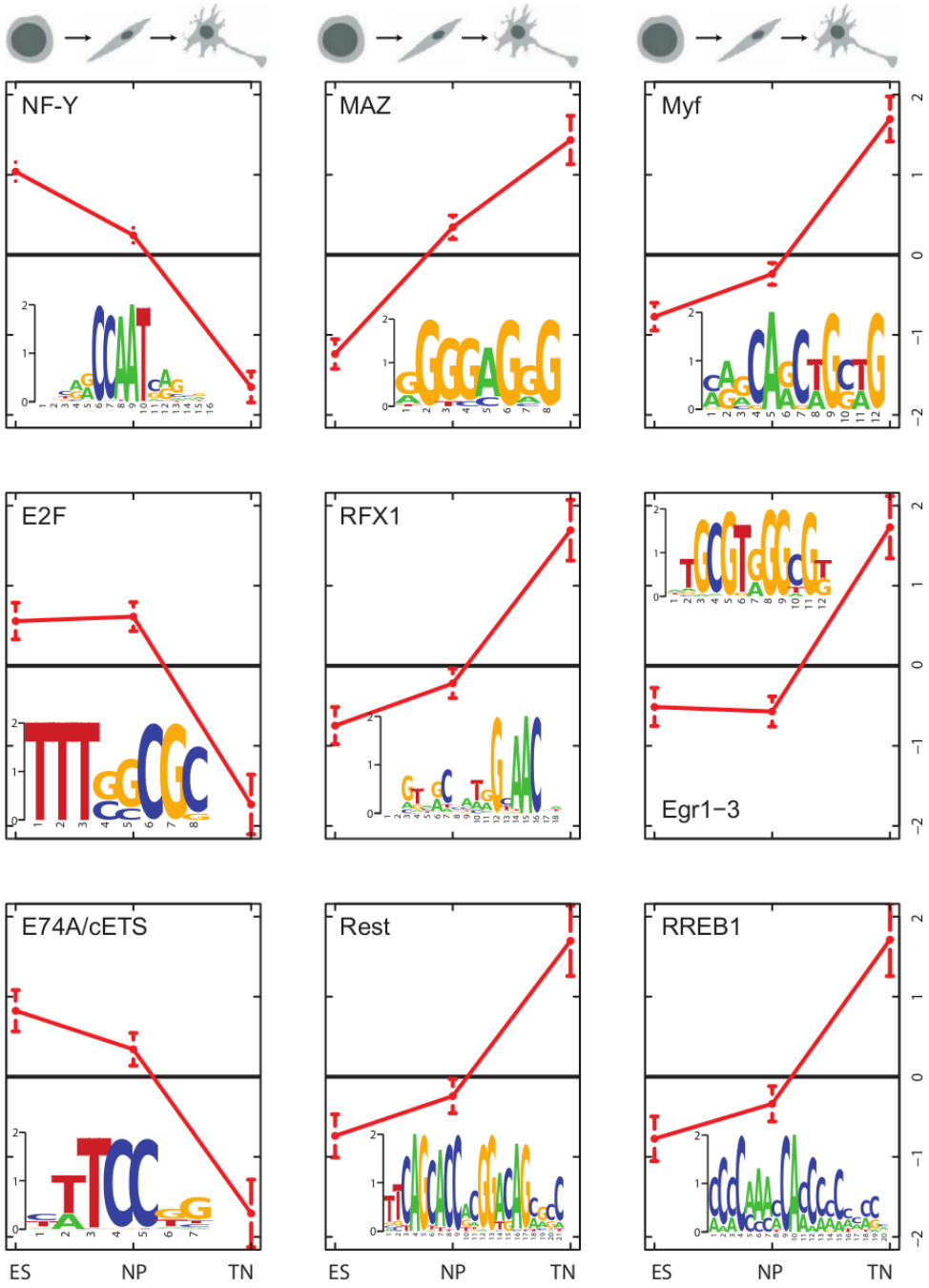
17. Chan, H.M. and La Thangue, N.B. 2001. p300/CBP proteins: HATs for transcriptional bridges and scaffolds. J Cell Sci 114(Pt 13): 2363-2373.
18. Chen, Z.F., Paquette, A.J., and Anderson, D.J. 1998. NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. Nat Genet 20(2): 136-142.
19. Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. 2002. Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell 111(2): 185-196.
20. Das, D., Nahle, Z., and Zhang, M.Q. 2006. Adaptively inferring human transcriptional subnetworks. Mol Syst Biol 2: 2006 0029.
21. de Hoon, M. and Hayashizaki, Y. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques 44(5): 627-628, 630, 632.
22. Dietrich, N., Lerdrup, M., Landt, E., Agrawal-Singh, S., Bak, M., Tommerup, N., Rappsilber, J., Sodersten, E., and Hansen, K. 2012. REST-Mediated Recruitment of Polycomb Repressor Complexes in Mammalian Cells. PLoS Genet 8(3): e1002494.
23. Enderle, D., Beisel, C., Stadler, M.B., Gerstung, M., Athri, P., and Paro, R. 2010. Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. Genome Res.
24. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473(7345): 43-49.
25. Feng, Y.Q., Seibler, J., Alami, R., Eisen, A., Westerman, K.A., Leboulch, P., Fiering, S., and Bouhassira, E.E. 1999. Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange. J Mol Biol 292(4): 779-785.
26. Gao, F., Foat, B.C., and Bussemaker, H.J. 2004. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics 5: 31.
27. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10): R80.
28. Harbers, M. and Carninci, P. 2005. Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2(7): 495-502.
29. Herranz, N., Pasini, D., Diaz, V.M., Franci, C., Gutierrez, A., Dave, N., Escriva, M., Hernandez-Munoz, I., Di Croce, L., Helin, K. et al. 2008. Polycomb complex 2 is required for E-cadherin repression by the Snail1 transcription factor. Mol Cell Biol 28(15): 4772-4781.
30. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830): 1497-1502.
31. Johnson, R., Teh, C.H., Kunarso, G., Wong, K.Y., Srinivasan, G., Cooper, M.L., Volta, M., Chan, S.S., Lipovich, L., Pollard, S.M. et al. 2008. REST regulates distinct transcriptional networks in embryonic and neural stem cells. PLoS Biol 6(10): e256.
32. Jorgensen, H.F., Chen, Z.F., Merkenschlager, M., and Fisher, A.G. 2009. Is REST required for ESC pluripotency? Nature 457(7233): E4-5; discussion E7.
33. Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P. et al. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 17(6): 691-707.

34. Kouzarides, T. 2007. Chromatin modifications and their function. Cell 128(4): 693-705.
35. Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S. et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet 4(10): e1000242.
36. Landeira, D., Sauer, S., Poot, R., Dvorkina, M., Mazzarella, L., Jorgensen, H.F., Pereira, C.F., Leleu, M., Piccolo, F.M., Spivakov, M. et al. 2010. Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. Nat Cell Biol 12(6): 618-624.
37. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3): R25.
38. Lee, M.G., Wynder, C., Cooch, N., and Shiekhattar, R. 2005. An essential role for CoREST in nucleosomal histone 3 lysine 4 demethylation. Nature 437(7057): 432-435.
39. Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schubeler, D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet.
40. Liu, Y., El-Naggar, S., Darling, D.S., Higashi, Y., and Dean, D.C. 2008. Zeb1 links epithelial-mesenchymal transition and cellular senescence. Development 135(3): 579-588.
41. Liu, Y., Shao, Z., and Yuan, G.C. 2010. Prediction of Polycomb target genes in mouse embryonic stem cells. Genomics 96(1): 17-26.
42. Lynch, M.D., Smith, A.J., De Gobbi, M., Flenley, M., Hughes, J.R., Vernimmen, D., Ayyub, H., Sharpe, J.A., Sloane-Stanley, J.A., Sutherland, L. et al. 2011. An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J 31(2): 317-329.
43. Margueron, R. and Reinberg, D. 2011. The Polycomb complex PRC2 and its mark in life. Nature 469(7330): 343-349.
44. Meissner, A. 2010. Epigenetic modifications in pluripotent and differentiated cells. Nat Biotechnol 28(10): 1079-1088.
45. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B. et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454(7205): 766-770.
46. Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. 2010. GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet 6(12): e1001244.
47. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. et al 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153): 553-560.
48. Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schubeler, D. 2008. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30(6): 755-766.

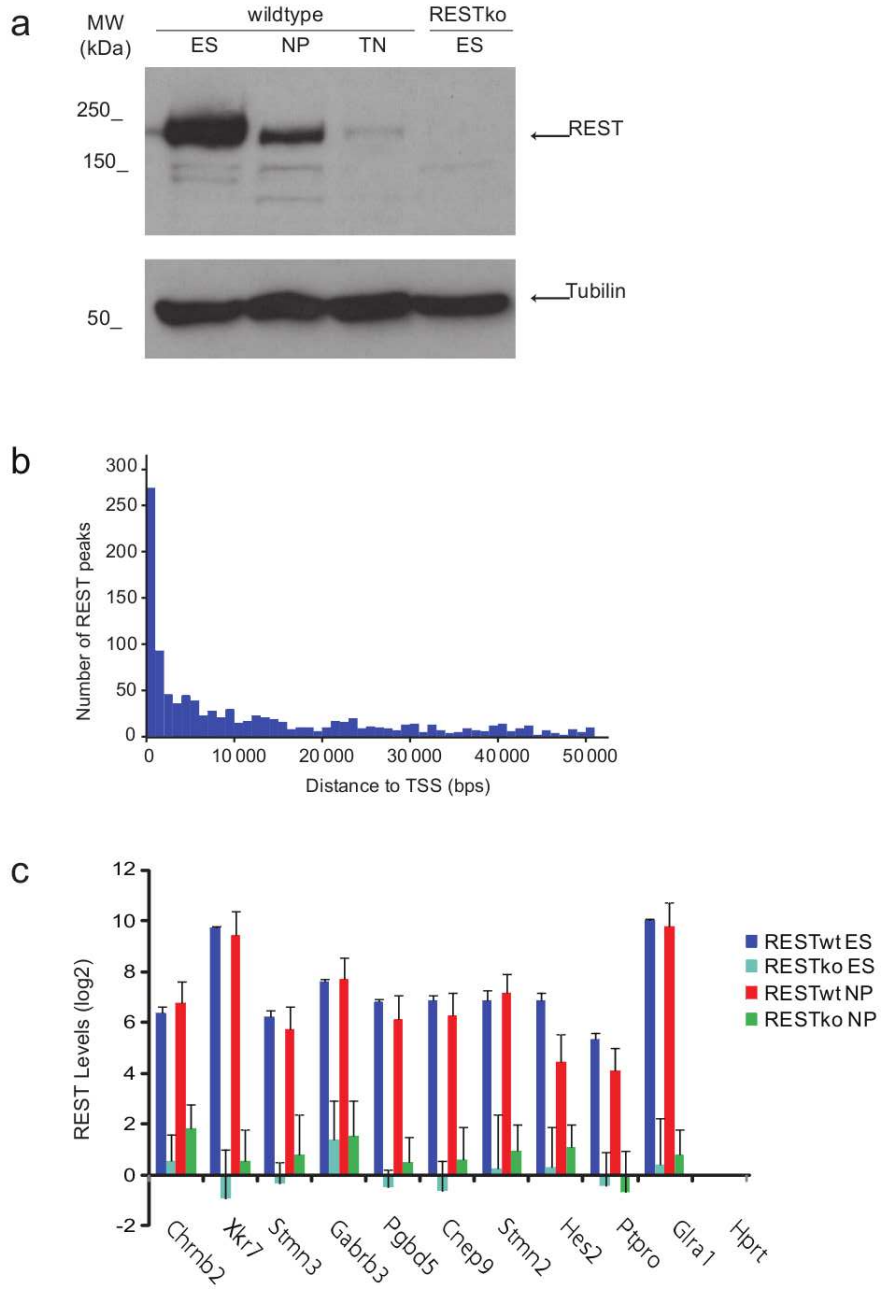
49. Muller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. 2002. Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell 111(2): 197-208.
50. Nieto, M.A. 2002. The snail superfamily of zinc-finger transcription factors. Nat Rev Mol Cell Biol 3(3): 155-166.
51. O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. 2001. The polycomb-group gene Ezh2 is required for early mouse development. Mol Cell Biol 21(13): 4330-4336.
52. Pasini, D., Cloos, P.A., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. 2010. JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature 464(7286): 306-310.
53. Pereira, C.F., Piccolo, F.M., Tsubouchi, T., Sauer, S., Ryan, N.K., Bruno, L., Landeira, D., Santos, J., Banito, A., Gil, J., Koseki, H. et al. 2010. ESCs require PRC2 to direct the successful reprogramming of differentiated cells toward pluripotency. Cell Stem Cell 6(6): 547-556.
54. Plachta, N., Bibel, M., Tucker, K.L., and Barde, Y.A. 2004. Developmental potential of defined neural progenitors derived from mouse embryonic stem cells. Development 131(21): 5449-5456.
55. Postigo, A.A. and Dean, D.C. 2000. Differential expression and function of members of the zfh-1 family of zinc finger/homeodomain repressors. Proc Natl Acad Sci U S A 97(12): 6391-6396.
56. Ren, X. and Kerppola, T.K. 2011. REST interacts with Cbx proteins and regulates polycomb repressive complex 1 occupancy at RE1 elements. Mol Cell Biol 31(10): 2100-2110.
57. Ringrose, L. and Paro, R. 2007. Polycomb/Trithorax response elements and epigenetic memory of cell identity. Development 134(2): 223-232.
58. Satterlee, J.S., Schubeler, D., and Ng, H.H. 2010. Tackling the epigenome: challenges and opportunities for collaboration. Nat Biotechnol 28(10): 1039-1044.
59. Schoenherr, C.J. and Anderson, D.J. 1995. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. Science 267(5202): 1360-1363.
60. Schuettengruber, B. and Cavalli, G. 2009. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development 136(21): 3531-3542.
61. Schwartz, Y.B. and Pirrotta, V. 2008. Polycomb complexes and epigenetic states. Curr Opin Cell Biol 20(3): 266-273.
62. Simon, J.A. and Kingston, R.E. 2009. Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol 10(10): 697-708.
63. Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K. et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480(7378): 490-495.
64. Sun, Y.M., Greenway, D.J., Johnson, R., Street, M., Belyaev, N.D., Deuchars, J., Bee, T., Wilde, S., and Buckley, N.J. 2005. Distinct profiles of REST interactions with its target genes at different stages of neuronal development. Mol Biol Cell 16(12): 5630-5638.

65. Suzuki, H. Forrest, A.R. van Nimwegen, E. Daub, C.O. Balwierz, P.J. Irvine, K.M. Lassmann, T. Ravasi, T. Hasegawa, Y. de Hoon et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet 41(5): 553-562.
66. Suzuki, M.M. and Bird, A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9(6): 465-476.
67. Tao, Y., Kassatly, R.F., Cress, W.D., and Horowitz, J.M. 1997. Subunit composition determines E2F DNA-binding site specificity. Mol Cell Biol 17(12): 6994-7007.
68. Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. Science 329(5992): 689-693.
69. van Nimwegen, E. 2007. Finding regulatory elements and regulatory motifs: a general probabilistic framework. BMC Bioinformatics 8 Suppl 6: S4.
70. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. 2009. A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10(4): 252-263.
71. Vlieghe, D., Sandelin, A., De Bleser, P.J., Vleminckx, K., Wasserman, W.W., van Roy, F., and Lenhard, B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res 34(Database issue): D95-97.
72. Wasserman, W.W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5(4): 276-287.
73. Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39(4): 457-466.
74. Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24(1): 238-241.
75. Yamada, Y., Aoki, H., Kunisada, T., and Hara, A. 2010. Rest promotes the early differentiation of mouse ESCs but is not required for their maintenance. Cell Stem Cell 6(1): 10-15.
76. Zhang, C., Xuan, Z., Otto, S., Hover, J.R., McCorkle, S.R., Mandel, G., and Zhang, M.Q. 2006. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res 34(8): 2238-2246.
77. Zheng, D., Zhao, K., and Mehler, M.F. 2009. Profiling RE1/REST-mediated histone modifications in the human genome. Genome Biol 10(1): R9.
78. Zhou, V.W., Goren, A., and Bernstein, B.E. 2011. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12(1): 7-18.

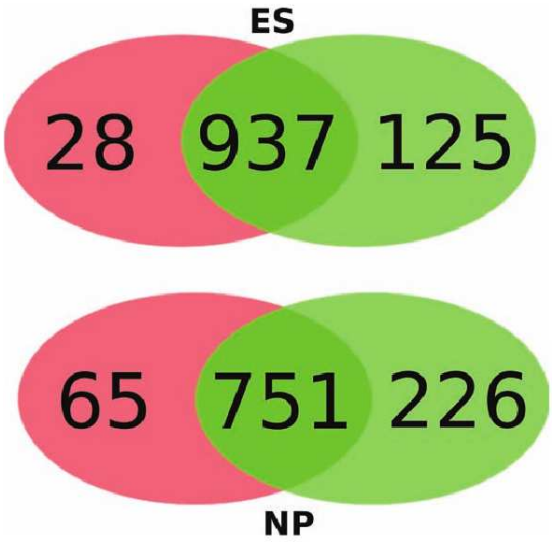
2.7 Supplementary Figures



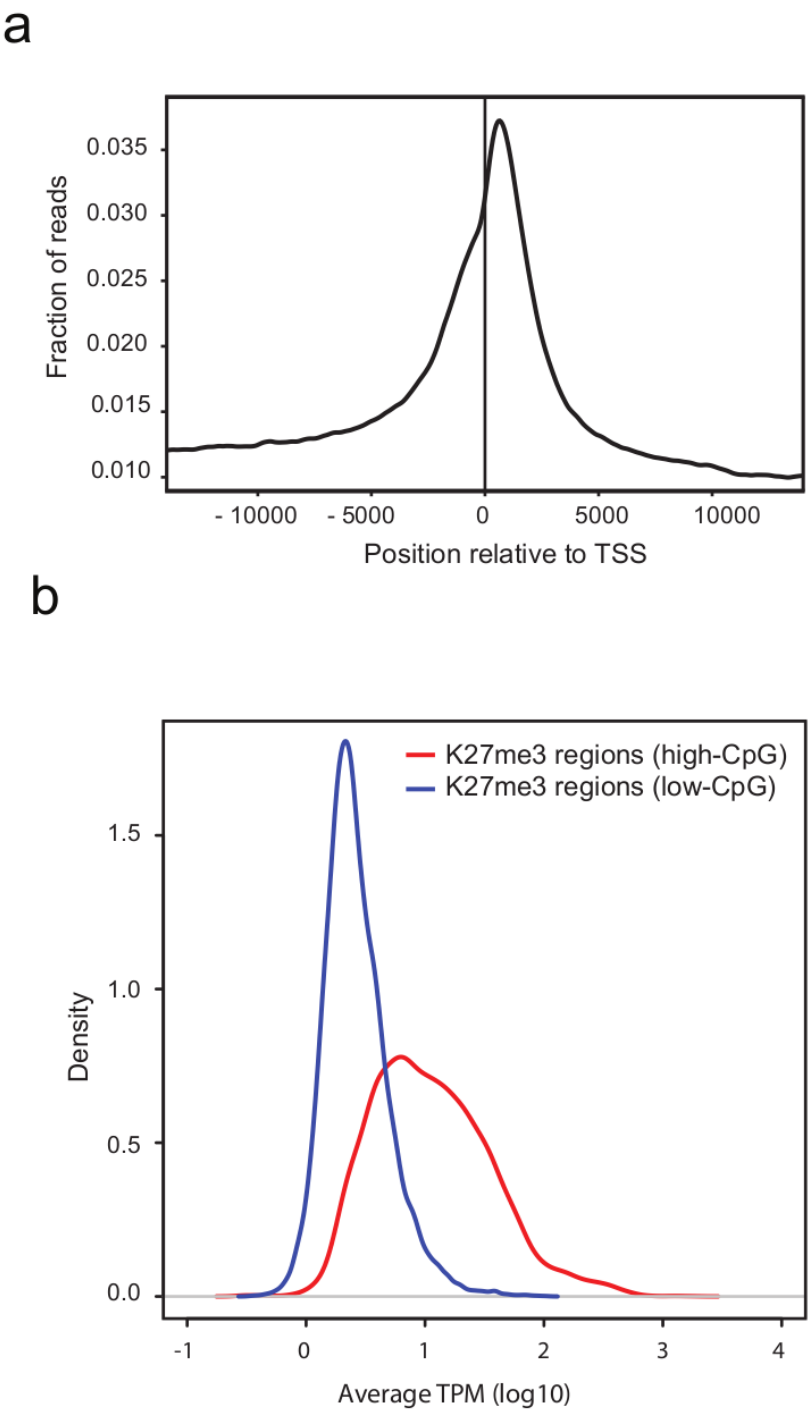
Supplementary Figure 2.1: MARA expression analysis predicts transcription factors explaining transcriptional dynamics during neuronal differentiation: Normalized activity profiles of the nine most significant motifs that explain changes in gene expression during the differentiation process (red lines, with standard errors indicated). The three time points correspond to the ES, NP and TN stage. Sequence logos of each of the motifs and the transcription factors thought to bind to them are shown as insets.



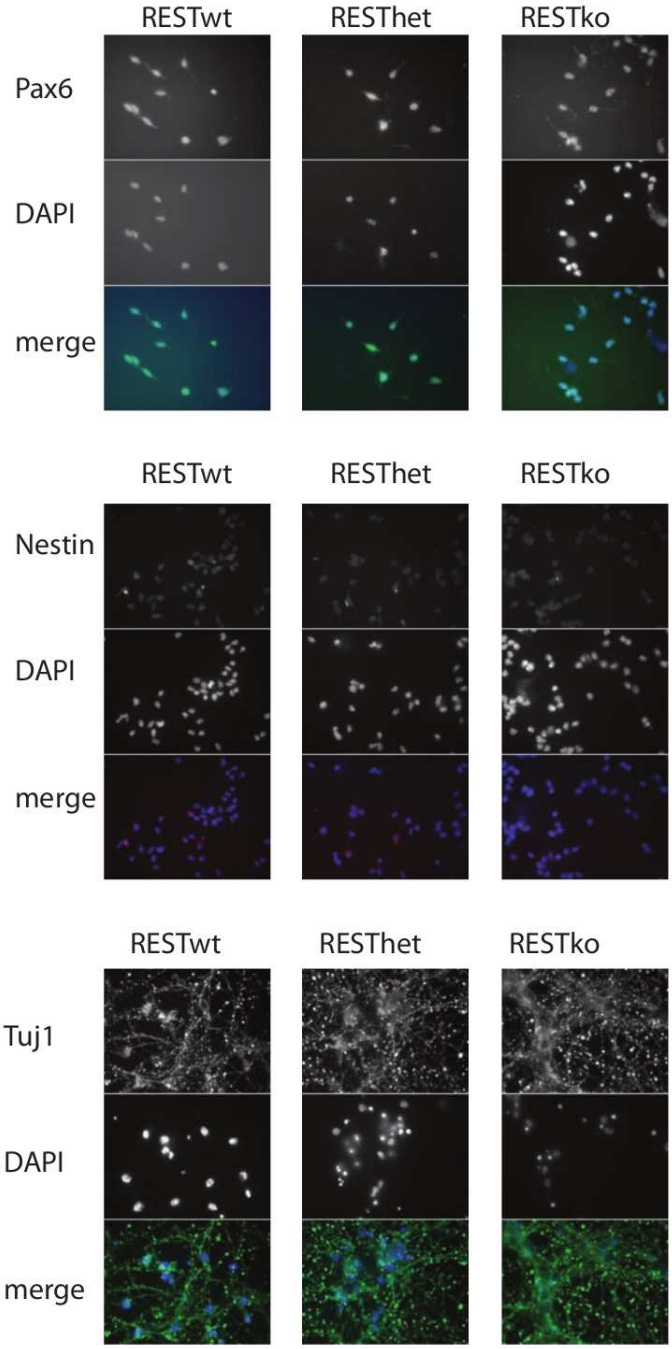
Supplementary Figure 2.2: REST expression and binding during differentiation: **a)** Protein levels of REST as detected by Western Blot in extracts from ES, NP and TN in wildtype as well as RESTko background (upper panel). Tubulin serves as a loading control (lower panel). **b)** Distribution of the distance between REST binding peaks and the nearest transcription start site (TSS). **c)** Quantitative PCR of REST ChIPs at the ES and NP stage confirms all 10 tested sites of REST binding as identified by ChIP-seq. Enrichments are normalized to a negative control (Hprt). RESTko cells show no signal confirming the specificity of the antibody. Shown are mean enrichments. Error bars show the standard deviation of three biological replicates.



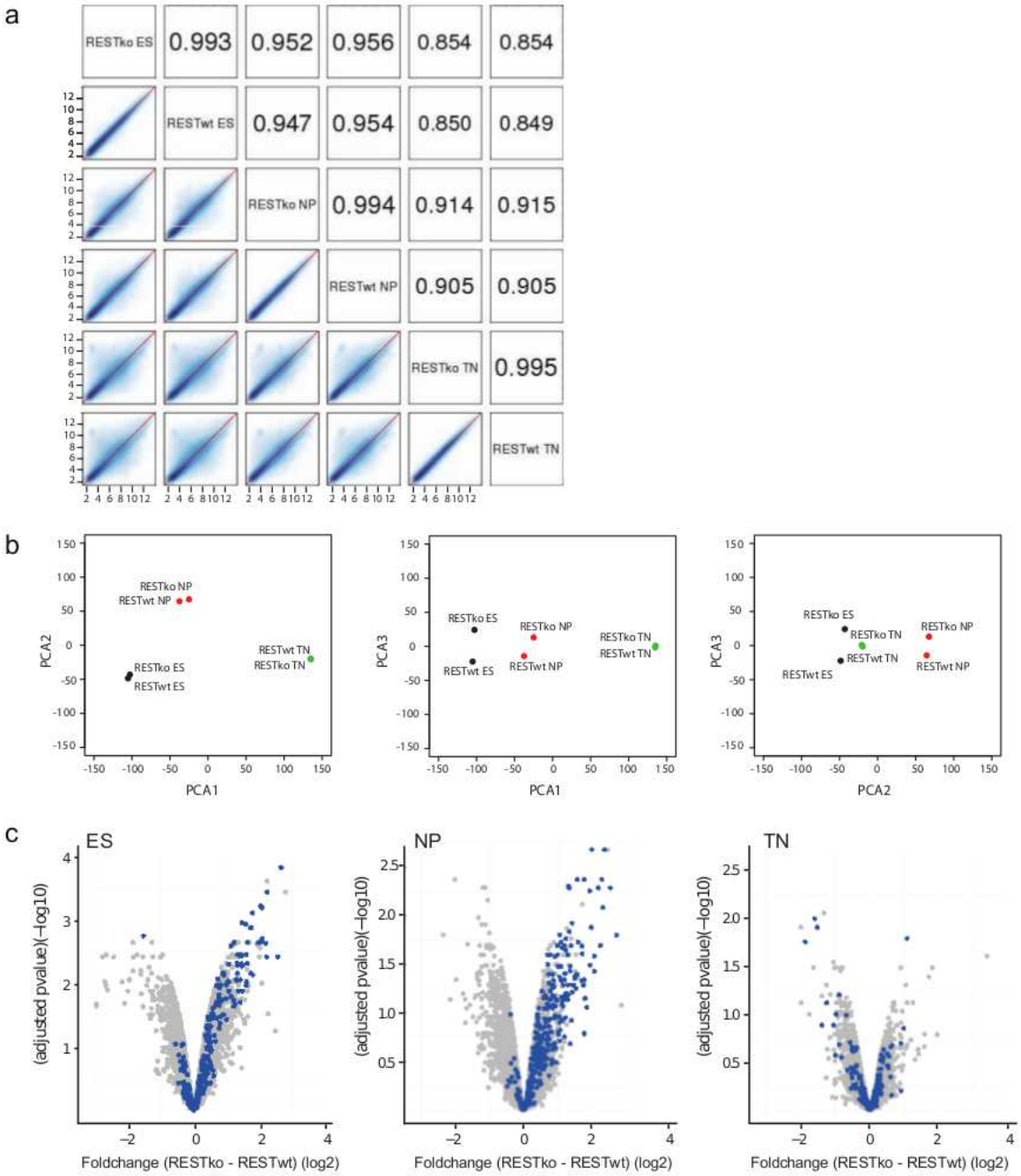
Supplementary Figure 2.3: Comparison of the REST binding sites identified in our study with those identified by Johnson et al. 2. We calculated REST enrichments from our ChIP-seq data for all genomic regions reported by Johnson et al. as binding sites of REST at either the ES or NP stage. Next, we calculated what fraction of regions passed the cut-off of an enrichment of 2 according to either data-set. This is shown separately in a Venn diagram (red corresponds to Johnson et al., green corresponds to our ChIP-seq data) for the ES (upper panel) and NP (lower panel) stage. Note that while both studies use ES cells as starting point distinct differentiation protocols are used leading to different neuronal populations, which explains larger variation in the differentiated state.



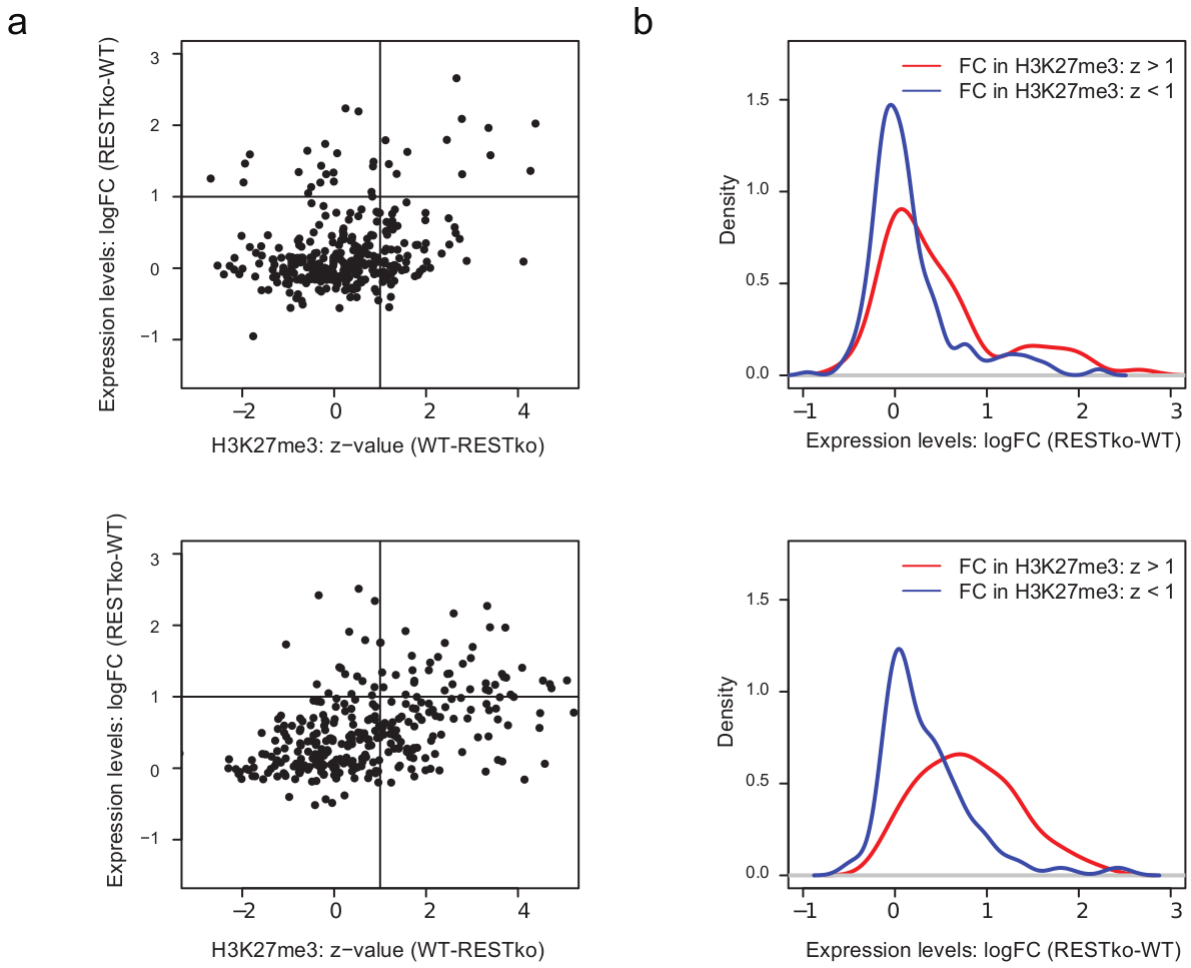
Supplementary Figure 2.4: **a)** H3K27me3 signal peaks downstream of transcription start sites: Spatial distribution of H3K27me3 signal relative to TSS. All H3K27me3 ChIP-seq samples were pooled and the average read density was plotted relative to position around TSS (14000bp). This reveals a H3K27me3 peak in the first 1000bps downstream of TSS. **b)** Distributions of the absolute H3K27me3 levels (log ChIP-seq reads per million averaged across the three stages) at all high-CpG (red) and low-CpG regions (blue) that are significantly enriched for H3K27me3.



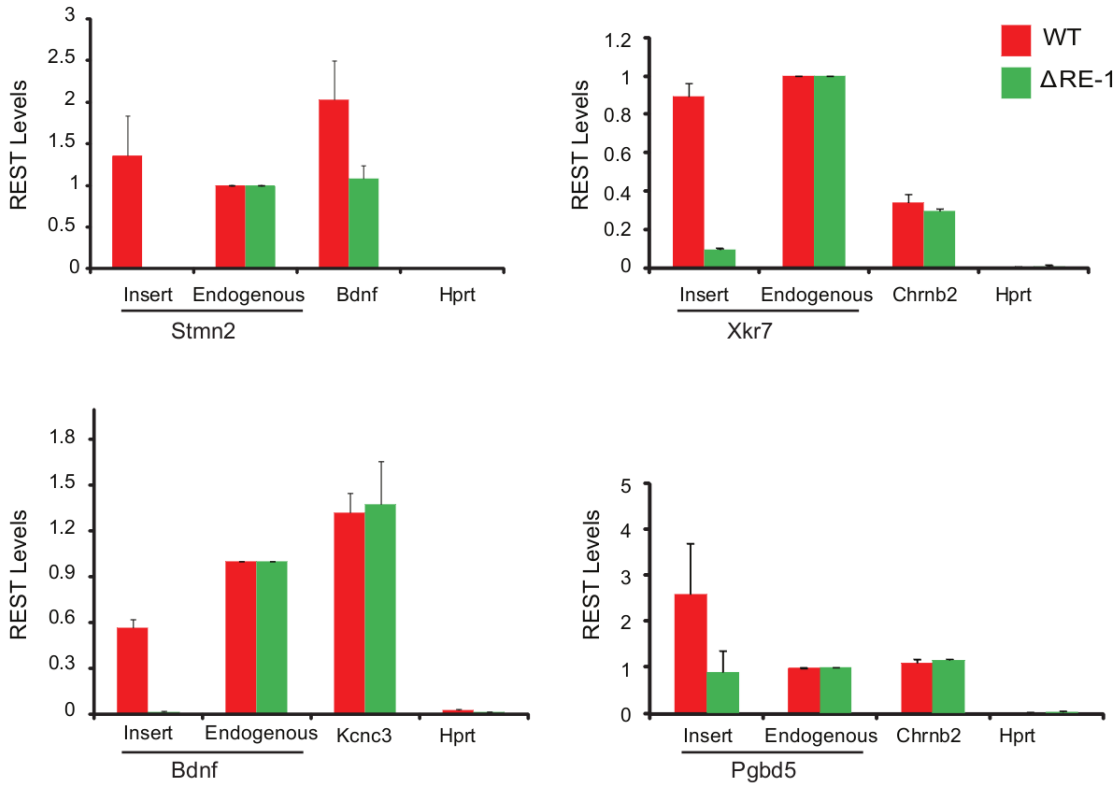
Supplementary Figure 2.5: REST knockout ES cells form neurons similar to wildtype ES cells PartI: Marker proteins show similar staining patterns in immuno-cytochemistry in REST knockout and wildtype (WT) cells: REST wildtype (RESTwt), heterozygous (RESThet) and homozygous knockout (RESTko) neuronal progenitors and terminal neurons were fixed and stained for several marker proteins specific for the neuronal progenitor stage (Pax6 (top panel) and Nestin (middle panel)) and terminal neuron stage (Tuj1 (bottom panel), respectively. The cells shown are representative for the population.



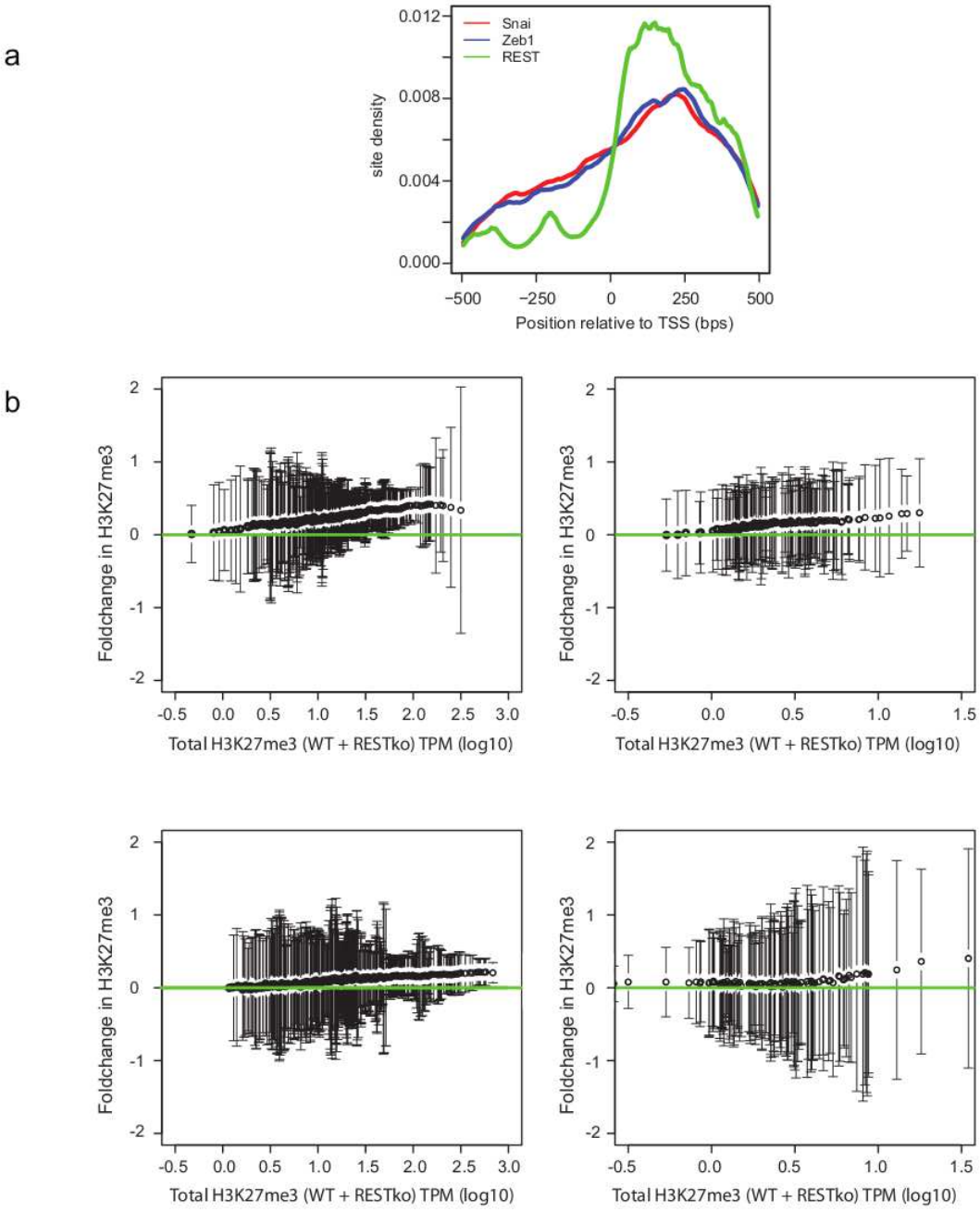
Supplementary Figure 2.6: REST knockout ES cells form neurons similar to wildtype ES cells PartII: **a)** Pairwise correlations of all gene expression microarrays. Shown are the normalized mean expression levels of three biological replicates each. RESTko cells show the highest correlation to each corresponding REST wildtype stage illustrating that only a small number of genes is affected in gene expression. RESTko and REST wildtype terminal neurons show the highest correlation. **b)** Principal component analysis shows that RESTko cells cluster with the corresponding wildtype stage. c) Volcano plots depict the fold-change in gene expression in RESTko vs RESTwt cells and the corresponding adjusted p-value for all three stages of differentiation. REST target genes are colored in blue.



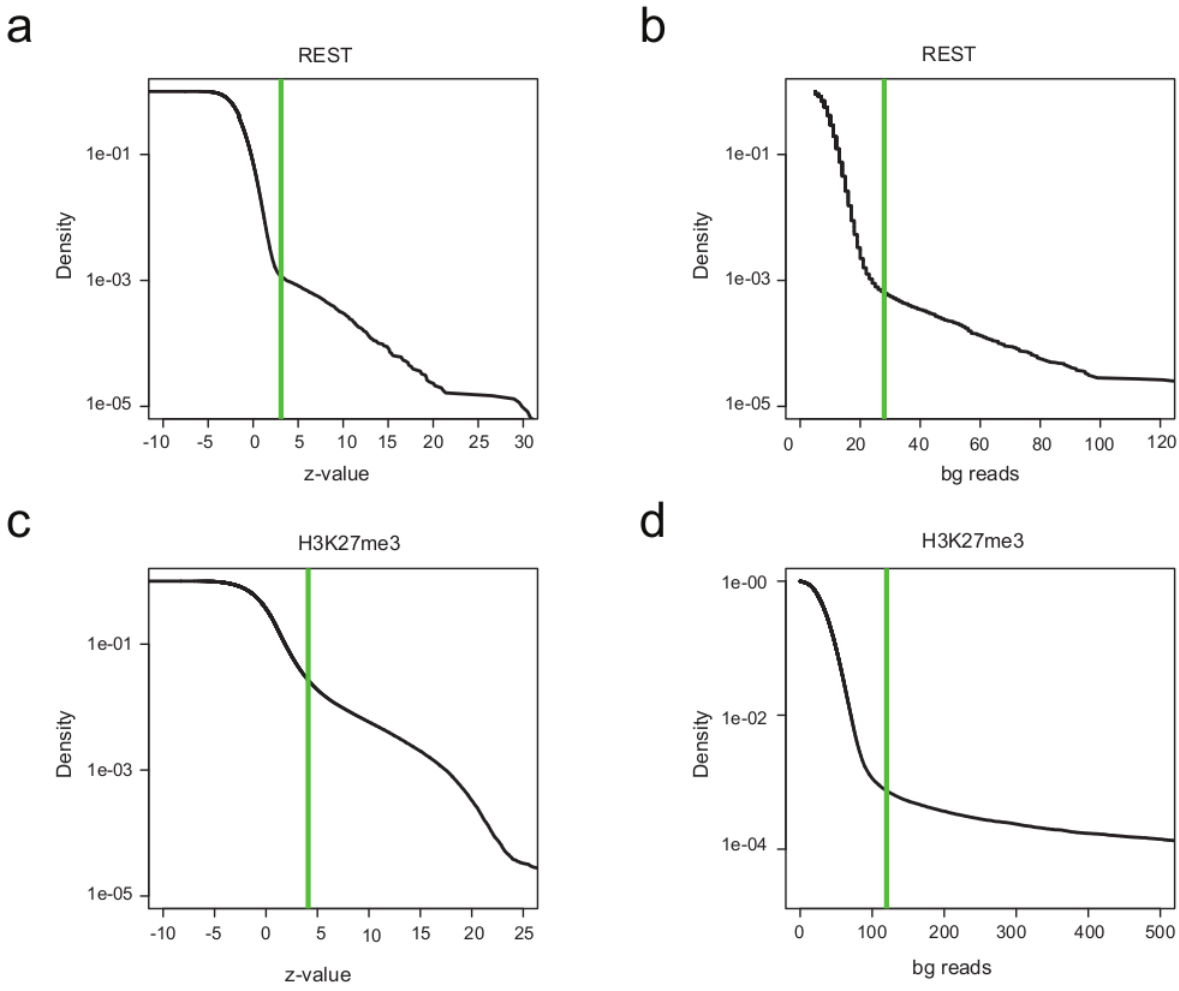
Supplementary Figure 2.7: Independent changes in H3K27me3 levels and gene expression levels at many REST targets. **a)** Pairwise comparison of changes in the significance of H3K27me3 levels (z-value wildtype Minus RESTko, horizontal axis) and changes in transcription (log fold-change RESTko minus wildtype, vertical axis) at both the ES (top panel) and NP stages (bottom panel). The horizontal and vertical lines correspond to a z-value of 1 and a log fold-change of 1 (i.e. two-fold upregulation). In general there is only a weak correlation between the amount of H3K27me3 loss and transcriptional up-regulation. At the ES stage many REST targets are transcriptionally up-regulated without showing a loss of H3K27me3. At the NP stage a significant fraction (33%) of REST targets that significantly lose H3K27me3 (z¿1) are not significantly up-regulated. **b)** Distribution of expression log fold-changes under RESTko for REST targets that significantly lose H3K27me3 (z¿1, red lines) and REST targets that do not significantly lose H3K27me3 (z¡1, blue lines), both at the ES (top panel) and NP (bottom panel) stages. As expected there is an overall association between loss of H3K27me3 and transcriptional up-regulation, especially at the NP stage.



Supplementary Figure 2.8: RE-1 binding site is essential for REST binding. Transgenic wildtype promoters show strong REST binding, but no or weak binding at the four RE-1 mutant sequences. Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. All REST levels are scaled to that of the endogenous region and error-bars show the standard error of three biological replicates.



Supplementary Figure 2.9: **a)** Distributions of predicted Snail (red), Zeb-1 (blue) and REST (green) binding sites are located downstream of transcription start sites. Shown is the frequency of predicted binding. **b)** Shown are on the x-axis the total H3K27me3 levels (sum of wildtype and RESTko signal) and on the y-axis the average H3K27me3 fold-changes between wildtype and RESTko with standard errors (black dots with error-bars) for all non REST target regions separated into high-CpG (left) and low-CpG (right) regions and for both the ES (top) and NP (bottom) stages.



Supplementary Figure 2.10: **a)** Reverse-cumulative distribution of the z-statistic for enrichment of ChIP-seq reads from the REST IPs relative to background for 1kb windows genome-wide. The distribution clearly shows two regions with a second tail at z-statistics larger than approximately 3. The vertical line ($z = 3.1$) shows the cut-off that we chose for considering a window significantly enriched for REST binding. The cut-off was chosen so as to ensure good sensitivity in REST binding region identification. **b)** Reverse-cumulative distribution of the number of background reads per 1kb window genome-wide. The vertical axis is shown on a logarithmic scale. We observe that the distribution drops steeply up to approximately 20 reads per window, after which the distribution shows a long tail with some windows showing over 100 reads. We filter these genomic regions with aberrantly high background counts (vertical line). **c)** Reverse-cumulative distribution of the z-statistic for enrichment of ChIP-seq reads from the H3K27me3 IPs relative to background for 2kb windows genome-wide. Again two (and maybe even three) regimes in the distribution are clearly evident and we chose a cut-off of $z = 4.0$ (vertical line) to identify windows significantly enriched for H3K27me3. **d)** Reverse-cumulative distribution of the number of background reads per 2kb window genome-wide. The vertical axis is shown on a logarithmic scale. We observe that the distribution drops steeply up to approximately 100 reads per window, after which the distribution shows a long tail with some windows showing over 500 reads. We filter these genomic regions with aberrantly high background counts (vertical line).

2.8 Supplementary Tables

Stage	z-value for H3K27me3 enrichments
Embryonic stem cell	4.2
Neural progenitor	5.7
Terminal neuron	1.3

Supplementary Table 2.1: For each of the three stages we compared the mean and variance of the H3K27me3 signal in all REST target promoters (predicted number of binding sites larger than 0.75) and all non-target promoters and calculated a z-value for the difference. The positive numbers in all three stages show that REST target promoters show more H3K27me3 than non-target promoters at all stages.

Promoter class/target	REST target	Non-target
CpG-high	405 (1.8%)	12886 (56.9%)
CpG-low	93 (0.4%)	9255 (40.1%)

Supplementary Table 2.2: Numbers (and corresponding percentages) of promoters in high- and low-CpG classes that are either targeted by REST (i.e. with a REST binding peak within 2kb of the TSS) or that are not a REST target. The numbers show that REST predominantly targets high-CpG promoters, as high-CpG promoters are roughly 3 times more likely to be REST targets than low-CpG promoters.

H3K27me3 region class/target	REST target	Non-target
Proximal	351 (1.9%)	5250 (28.7%)
Distal	199 (1.1%)	12496(68.3%)

Supplementary Table 2.3: Numbers (and corresponding percentages) of H3K27me3-enriched regions that are either proximal or distal to a TSS, and that are either targeted by REST or not (Non-target). REST predominantly targets proximal regions with H3K27me3, as proximal regions are 4 times more likely to be a REST target than distal regions.

Observables	H3K27me3 levels (ChIP-chip)	Transcript levels (chip)
Fraction of explained variance	7.8%	6.9%

Supplementary Table 2.4: For each promoter p we calculate the variance of both its expression levels (as measured by micro-array probes for the associated transcripts and its H3K27me3 levels (as measured by ChIP-chip).By comparing the observed expression and H3K27me3 levels, with the levels predicted by MARA[1] and Epi-MARA using the linear model with the fitted motif activities A_{ms}, we can calculate what fraction of the total variance, i.e. summed over all promoters, is explained by the linear model. The table shows that the linear model captures a similar fraction of variance of both the H3K27me3 and transcript level dynamics across the differentiation.

2.9 Supplementary Methods

2.9.1 Epi-MARA

Epi-MARA models the dynamics of epigenetic marks in terms of predicted TFBSs in regulatory regions genome-wide, building on the Motif Activity Response Analysis that we developed previously[1].

Transcription factor binding site predictions for promoters

For the Epi-MARA analysis of the ChIP-chip data, we selected all promoters for which we had H3K27me3 ChIP-chip measurements. Proximal promoter regions were constructed by taking the transcription start sites obtained from UCSC (RefSeq IDs) and extending them by ± 500bps. We have established in previous computational analyses that most functional TFBSs occur in these areas[3]. For each proximal promoter region we used the UCSC pairwise alignment to extract orthologous sequences from mouse, human, rhesus macaque, dog, cow, horse, and opossum. We then used T-Coffee[4] to create multiple alignments of the orthologous proximal promoter sequences. Using databases of experimentally determined binding sites[5,6], we curated a set of 207 mammalian regulatory motifs (position specific weight matrices) representing the binding specificities of approximately 350 mammalian TFs. The curation methodology was described previously[1] and involves removal of mutually redundant motifs and associating motifs with their respective binding factors by comparing the protein sequences of DNA binding domains of TFs.
To predict transcription factor binding sites on our multiple alignments we use our MotEvo algorithm[7]. MotEvo is a Bayesian probabilistic method that treats the alignments as a mixture of columns that are evolving neutrally, segments that are binding sites for one of the motifs that are evolving under the constraints set by the requirement that the sequence segments remain their affinity for the cognate TF, and segments that are under purifying selection of unknown function. Besides the multiple alignments and position specific weight matrices, MotEvo also takes the phylogenetic tree of the species as input, which we obtained by comparing the third positions of fourfold degenerate codons of orthologous proteins. MotEvo then assigns, to each position in the multiple alignments, a posterior probability that a binding site for each of the motifs in our collection occurs at this position. Finally, MotEvo estimates, separately for each motif, the distribution of binding site occurrences as a function of position relative to TSS, and updates the posterior probabilities of predicted sites by taking these positional preferences into account. Finally, we summarize the TFBS predictions by a matrix with components N_{pm}, denoting the sum of the posterior probabilities of all binding sites for motif m in promoter p.

Quantifying H3K27me3 levels

For the ChIP-chip data, each probe measures the ratio of immunoprecipitated fragments and untreated (Input) fragments. The logarithms of these ratios are normalized using standard procedures for oligonucleotide micro-arrays. The H3K27me3 signal at a given promoter is obtained by averaging the log-ratios of the probes intersecting the promoter. We additionally average the signal over biological replicate experiments. For the ChIP-seq data, we quantity the H3K27me3 at a promoters by collecting reads that overlap a 4kb region centered at the transcription start site. The 4kb length was chosen based on our analysis of H3K27me3 regions genome-wide. We find that the majority of H3K27me3 enriched regions are 3-4kb in length (Supplementary Fig. 2.4a). Thus, by summing reads over a 4kb region, we generally collect reads from the entire H3K27me3 enriched region (thereby reducing fluctuations) while on the other hand not diluting the signal by including flanking regions that are not enriched for H3K27me3. To reduce fluctuations for regions that have low read counts we add a pseudo-count to the read count of each 4kb region. The size of the pseudo-count was chosen to be the average number of reads per region in the background (Input) sample. Finally, we normalize read-counts by the total number of reads in the sample and take the logarithms to obtain the final quantification

M_{pt} of the occurrence of the epigenetic mark by at promoter p at time t. In addition, we average the level M_{pt} over available biological replicates.

Fitting the linear model

We then model the chromatin mark levels across time in terms of the predicted binding sites using following linear model:

$$M_{pt} = \text{noise} + c_p + k_t + \sum_m N_{pm} \cdot A_{mt} \tag{2.5}$$

where c_p is the basal level of the chromatin mark, k_t is a constant term accounting for the total expression at time t, and A_{mt} is the unknown activity of motif m at time point t. We assume the deviations between model and measured level M_{pt} (i.e. the 'noise' term in the above formula) is Gaussian distributed with the same, but unknown standard deviation σ for each time points. The likelihood of the chromatin mark signal then becomes

$$P(M|A,c,k,N,\sigma) \propto \prod_{pt} \frac{1}{\sigma} \exp\left[-\frac{\left(\sum_m N_{pm}A_{mt} - M_{pt} - c_p - k_t\right)^2}{2\sigma^2}\right]. \tag{2.6}$$

We first set the constants c_p and k_t to their maximal likelihood values, i.e. we maximize (2.5) with respect to c_p and k_t. This results in the following expression

$$P(M|A,N,\sigma) \propto \prod_{p,t} \exp\left[-\frac{\left(\sum_m \tilde{N}_{pm}\tilde{A}_{mt} - \tilde{M}_{pt}\right)^2}{2\sigma^2}\right], \tag{2.7}$$

where $\tilde{N}_{pm} = N_{pm} - < N_m >$ is the normalized matrix of site counts, i.e. where the average number of sites per promoter $< N_m >$ has been subtracted for each motif m, $\tilde{A}_{mt} = A_{mt} - < A_m >$ is normalized motif activity, i.e. where the average motif activity $< A_m >$ of motif m across the time course has been subtracted, and $\tilde{M}$ is the row and column normalized matrix of chromatin mark levels, i.e. where row and column averages have been subtracted from the matrix M such that all rows and columns of $\tilde{M}$ sum to zero.

To avoid over-fitting of the motif activities A_{mt}, we assign a Gaussian prior on each motif activity, i.e.:

$$P(\tilde{A}_{mt}|\lambda) \propto \prod_p \exp\left[-\frac{\lambda}{2\sigma^2}\left(\tilde{A}_{mt}\right)^2\right]. \tag{2.8}$$

Combining this prior with the likelihood (2.5) we obtain the following posterior probability distribution over the motif activities

$$P(A|M,N,\sigma) \propto \prod_{p,t} \frac{1}{\sigma} \exp\left[-\frac{\lambda\sum_m \left(\tilde{A}_{mt}\right)^2 + \left(\sum_m \tilde{N}_{pm}\tilde{A}_{mt} - \tilde{M}_{pt}\right)}{2\sigma^2}\right]. \tag{2.9}$$

In this expression we can analytically integrate over the unknown standard-deviation σ to obtain

$$P(A|M,N,\sigma) \propto \prod_t \exp\left[\frac{P\sum_{m,\tilde{m}}()\,W_{m\tilde{m}}()}{2\chi_t^2}\right], \tag{2.10}$$

where P is the total number of promoters, the matrix W is given by $W_{m\tilde{m}} = \sum_p \left(\tilde{N}_{pm}\tilde{N}_{p\tilde{m}} + \lambda\right)$, the $\tilde{A}^*_{mt}$ are the optimal motif activities, and the chi-squared deviation between model and measurements at time point

t is given by $\chi_t^2 = \sum_p \left(\tilde{M}_{pt} - \sum_m \tilde{N}_{pm} \tilde{A}_{mt} \right)^2$. The fitted motif activities are determined by singular variance decomposition of the matrix N_{pm}. Note that the resulting fitted activities will depend on the parameter λ of the Gaussian prior, i.e. the larger λ the smaller the inferred activities. In Epi-MARA, the variable λ of the Gaussian prior is chosen by a cross-validation procedure. We randomly select 80% of all promoters as a training set on which we fit the motif activities, and then evaluate the deviation between model and observed levels by M_{pt} on the 'test set' of the remaining 20% of promoters. The variable λ is chosen so as to minimize the error on the test set.

Finally, once λ has been determined, we infer both the maximal posterior activities A^*_{mt} and their standard-errors $\sigma_{mt} = \frac{\left(W^{-1}\right)_{m\tilde{m}} \chi_t^2}{P}$ from the multi-variant Gaussian posterior, in particular. To rank motifs by their importance in explaining variations in the levels M_{pt} we use a score similar to a z-statistic. The z-score z_m of motif m is quantified as an average squared z-value of the activity across conditions, i.e.

$$z_m = \sqrt{\frac{1}{T} \sum_t \left(\frac{\tilde{A}_{mt}}{\sigma_{mt}} \right)^2}, \tag{2.11}$$

where T is the number of time points in the data-set. Note that our z-scores are meant to rank the importance of motifs and cannot be used to assess the statistical significance of motif activities. To assess the statistical significance of a given z-score one can employ a standard randomization test. Using the same input data and site predictions, we randomize the association between promoters and site counts and run Epi-MARA, noting the resulting z-scores for all motifs. In this way we estimated that, for the H3K27me3 data we analyze, the probability of obtaining a z-score of 2.52 by chance is roughly $p = 5 \cdot 10^{-6}$.

Epi-MARA on H3K27me3 regions genome-wide

In order to run Epi-MARA on all H3K27me3 regions genome-wide we need to quantify H3K27me3 levels at each region across the time points. Different H3K27me3 regions have different sizes and to normalize H3K27me3 levels across all regions we first identified for each H3K27me3 region the 4kb bps window with the highest H3K27me3 levels. The H3K27me3 levels M_{rt} for each 4kb region r at each time point t were then obtained in the same way as for the promoters, i.e. we sum ChIP-seq reads intersecting the region, add a pseudo-count which corresponds to the average background level, divide by the total number of reads in the sample, and take the logarithm.

To obtain predicted site-counts N_{rm} for each region r, we first obtained multiple alignments for the entire 4kb regions as described above for the promoters. We then ran MotEvo to predict TFBSs for all motifs in the entire 4kb regions. To take into account the distinct base compositions for high-CpG and low-CpG regions, we used separate background models in MotEvo for the high-CpG and low-CpG regions by separately determining the frequencies of A, C, G, and T nucleotides in the high-CpG and low-CpG regions, resulting in:

H3K27me3	class A=T frequency	C=G frequency
high-CpG	0.197	0.303
low-CpG	0.269	0.231

Instead of taking the predicted binding sites from the entire 4kb regions for the site counts N_{rm}, we wanted to focus in on a 1kb window in each 4kb region that is most likely to contain the most relevant binding sites. We reasoned that the 1kb window with the highest overall density of predicted binding sites (which is typically the window that also shows the highest sequence conservation across mammals) is most likely to contain the relevant TFBSs and we determined N_{rm} using only the predicted TFBSs from this 1kb window.

Given that high-CpG and low-CpG H3K27me3 regions show distinct dynamics of H3K27me3 levels in general, we want to consider the possibility that the occurrences of TFBSs for a given motif m may have different effects at high-CpG and low-CpG regions. Thus we run Epi-MARA on all regions allowing for separate motif activities at high-CpG and low-CpG regions for all motifs. To infer motif activities separately for high- and low-CpG regions we treat, for each motif m, sites within low-CpG regions and sites within high-CpG regions as if they derived from two separate motifs, effectively doubling the number of motifs for which we infer activities. To do this we replace that original matrix of site counts with a new matrix N_{rm} that looks as follows:

$$N_{rm} = \begin{pmatrix} N_{rm}^{\text{high}} & 0 \\ 0 & N_{rm}^{\text{low}} \end{pmatrix}, \tag{2.12}$$

where $N_{rm}^{\text{high,low}}$ is the matrix of the number of expected sites using the predicted TFBSs for high-CpG and low-CpG regions, respectively. We then infer the activities in the same way as for the promoters.

To include the real binding data we have for the TF REST in the Epi-MARA analysis, we first transform the matrix N_{rm} into a binary matrix, that is to say all entries that are bigger than 0.2 (number of expected sites) are substituted by 1, otherwise by 0. The entries $N_{rm=REST}$ are then replaced by the binding data by putting a 1 if the H3K27me3 region p is a REST target (as defined earlier), and 0 if p is a non-target.

Determining H3K27me3 enriched regions genome-wide

To analyze H3K27me3 dynamics across the time course genome-wide, it is essentially that we obtain a reference set of regions for which to calculate H3K27me3 levels at each time point. To this end we decided to identify, genome-wide, all regions that are significantly enriched for H3K27me3 when considering all stages of the differentiation. We proceeded as follows. For each 2kb window on the genome, we calculate the fractions f_t of ChIP-seq reads from the samples at each time point t that map to the region in question as well as the fraction f_b of reads from the background sample that maps to the region. Assuming Poissonian noise in the fractions f_t , the variance of the fraction f_t is given by $V_t = \frac{f_t}{N_t}$, with N_t the total number of reads in the ChIP-seq sample at time t. The average fraction across the time course is simply given by $f = \frac{1}{T}\sum_t f_t$ and the variance associated with this fraction is $V = \frac{1}{T^2}\sum_t V_t$. Using this we obtain the following z-value for the overall enrichment at the region: $z = \frac{f_t - f_b}{\sqrt{V + \frac{f_b}{N_b}}}$. We plot the reverse-cumulative distribution of these z-values for all genomic windows of 2kb wide and chose a cut-off $z = 4.0$ to select significantly enriched regions (Supplementary Fig. 2.10).

2. MODELING OF EPIGENOME DYNAMICS IDENTIFIES TRANSCRIPTION FACTORS THAT MEDIATE POLYCOMB TARGETING

Chapter 3

MotEvo: Integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences

Phil Arnold, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen

published in *Bioinformatics*, December 15, 2011

Probabilistic approaches for inferring transcription factor binding sites (TFBSs) and regulatory motifs from DNA sequences have been developed for over two decades. Previous work has shown that prediction accuracy can be significantly improved by incorporating features such as the competition of multiple transcription factors (TFs) for binding to nearby sites, the tendency of TFBSs for co-regulated TFs to cluster and form cis-regulatory modules, and explicit evolutionary modeling of conservation of TFBSs across orthologous sequences. However, currently available tools only incorporate some of these features, and significant methodological hurdles hampered their synthesis into a single consistent probabilistic framework.

We present MotEvo, a integrated suite of Bayesian probabilistic methods for the prediction of TFBSs and inference of regulatory motifs from multiple alignments of phylogenetically related DNA sequences which incorporates all features just mentioned. In addition, MotEvo incorporates a novel model for detecting *unknown functional elements* that are under evolutionary constraint, and a new robust model for treating gain and loss of TFBSs along a phylogeny. Rigorous benchmarking tests on ChIP-seq datasets show that MotEvo's novel features significantly improve the accuracy of TFBS prediction, motif inference, and enhancer prediction.

3.1 Introduction

What the sequence specificities of different transcription factors (TFs) are and where in the genome their transcription factor binding sites (TFBSs) occur remain central questions in gene regulation. For over two

decades, a large number of computational methods has been developed that aim to support answering such questions, see e.g. [1,2] for reviews. Although much progress has been made, it remains highly challenging to obtain accurate computational TFBS predictions, especially on a genome-wide scale. For example, although from a biophysical point of view identical sequence segments should have equal affinity for the TF, one typically finds that only a small fraction of the sequences with high binding affinity act as functional TFBSs. That is, TFBS functionality is context-dependent and thus researchers have searched for additional features that are predictive for the functionality of putative sites.

One approach that has proven particularly fruitful is comparative genomic analysis of the conservation of putative TFBSs across related species, i.e. putative TFBSs that are highly conserved are generally more likely to be functional. A large number of approaches for incorporating conservation information has been proposed including several simple *ad hoc* methods, e.g. [3], but it has become clear that highest performance is obtained by methods that use explicit evolutionary models for the evolution of TFBSs along a phylogeny [4–6]. As a consequence, there has been considerable interest in extending methods for regulatory motif finding and TFBS prediction to include such explicit phylogenetic models. For example, the well-known Gibbs sampling [7] and Expectation-Maximization strategies [8] for *ab initio* motif finding, have been extended to methods that work on multiple alignments of orthologous sequences and use explicit evolutionary models [5, 9].

Beyond conservation information, other features have also proven highly useful in improving the accuracy of TFBS prediction. For example, especially in higher eukaryotes, functional TFBSs often come in clusters where multiple binding sites for a small subset of TFs co-occur in close proximity to each other [10]. Several methods were developed that, instead of looking for TFBSs for one TF at a time, explicitly look for clusters of sites for a collection of TFs. These methods have been especially successful in identifying cis-regulatory modules that are distal to their target gene [11, 12]. It has also proven useful to take into account the contribution of many weak binding sites and the competitive binding of multiple TFs to a DNA sequence by explicitly considering all possible configurations of non-overlapping binding sites using a dynamic programming procedure, e.g. [12–14].

However, currently these methodologies are spread over multiple computational tools that each only implement some of these methods. For example, the methods for finding cis-regulatory modules have been extended to analyze pairs of aligned species [15] but not to general multiple alignments and phylogenetic relationships. Other methods can only make predictions for one TF at a time, ignoring the competitive binding of multiple TFs, e.g. [4, 9], and the methods that incorporate sophisticated models for explicitly considering all possible binding configurations of multiple TFs cannot incorporate conservation information, e.g. [14]. Beyond this, as we show below, current methods that incorporate explicit evolutionary models make several implicit assumptions that cause 'pathologies' that significantly affect their performance.

Here we present a computational tool, MotEvo, that integrates a suite of Bayesian probabilistic methods for the prediction of regulatory sites and motifs on multiple alignments of phylogenetically related sequences. MotEvo not only implements and extends the functionality of many of the tools and methods mentioned in the introduction into a single integrated method, it also incorporates a number of new features that address the 'pathologies' that current methods suffer from, as we explicitly demonstrate below.

3.2 Methods

MotEvo takes as input either sequences from a single species or multiple alignments of orthologous sequences from several species, a collection of one or more position-specific weight matrices (WMs), and a phylogenetic tree relating the species. The user designates one of the species as the 'reference species' and MotEvo can then be asked to provide

- Posterior probabilities for a TFBS for each possible WM to occur at each position in the input sequences of the reference species.

Figure 3.1: A segment from a multiple alignment of orthologous mammalian sequences, together with an example configuration of 3 binding sites: one for the motif of the CCCTC-binding factor TF (CTCF) and two for the motif of the Serum Response Factor TF (SRF). Human is considered the reference species. Note that hypothesized TFBSs are not allowed to overlap within one configuration.

- The probability, at each position, for an *unknown functional element* (UFE) to occur, i.e. a TFBS for an unknown motif not contained in the input set.
- The estimated site-densities for each WM and for UFEs (possibly fitted to the data).
- Updated versions of the WMs (fitted to the data).
- Log-likelihood ratio scores, at each position, for the occurrence of cis-regulatory modules containing TFBSs for the input motifs.

We now discuss the methods that MotEvo uses to calculate these quantities.

3.2.1 Binding site configurations

We first introduce some notation. We denote by $\{S\}$ a collection of multiple alignments of sequences, by S an individual multiple alignment (or a segment from such an alignment), and by s an individual sequence or sequence segment. To indicate the segment of length l from sequence s, starting at position $(i+1)$ we use the notation $s_{[i,l]}$, and similarly $S_{[i,l]}$ indicates columns $(i+1)$ through $(i+l)$ of the multiple alignment S. Column numbers are always counted with respect to the position in the reference sequence. As in most approaches, we assume nucleotides at different positions in TFBSs are statistically independent and use position-specific weight matrices to represent TF binding specificities. We denote a collection of WMs by $\{w\}$ and a single WM from the set by w. The weight matrix entry w^i_α denotes the probability that nucleotide α occurs at position i of a binding site for WM w.

MotEvo considers all ways in which configurations of TFBSs (Fig. 3.1) for the WMs $\{w\}$ can be assigned to the sequences of the *reference species*. To explain how MotEvo calculates probabilities of possible configurations we first explain how MotEvo scores a single hypothesized TFBS.

3.2.2 Probabilities under the evolutionary model

Figure 3.2 shows a single hypothesized site for the CTCF motif from the TFBS configuration of Fig. 3.1. MotEvo calculates a probability ratio $P(S|w,T)/P(S|b,T)$ for observing this multiple alignment segment assuming that the sequences are evolving under constraints set by the WM w and assuming that the sequences are evolving 'neutrally' under a background model b, given the phylogenetic tree T. In contrast to most algorithms that implement explicit phylogenetic models, e.g. [4], MotEvo takes into account that functional TFBSs may only occur in a subset of the species. Sites may either have been truly lost or gained in some species during evolution, or sites may appear to have been lost as a consequence of errors in the multiple alignments. After experimenting with several procedures for treating these possibilities, including explicit

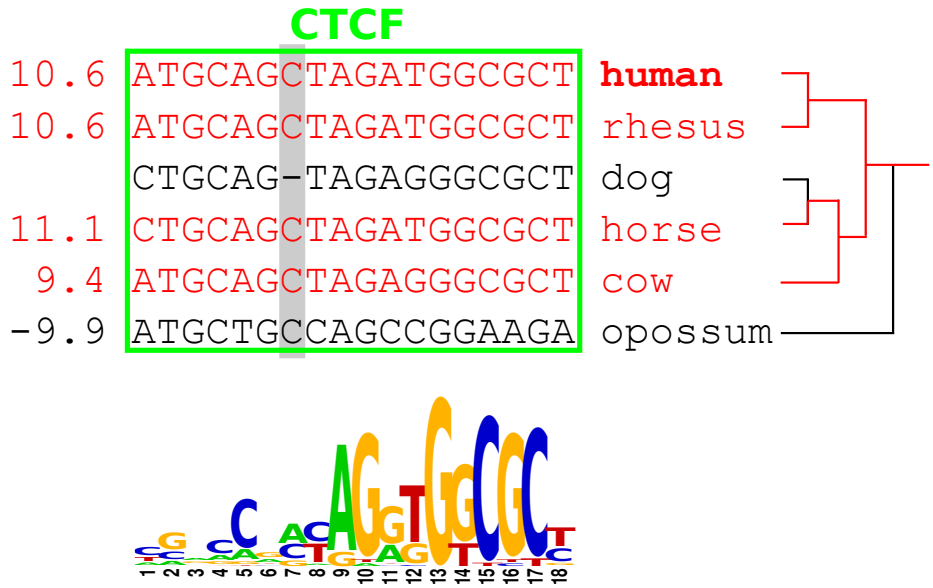


Figure 3.2: A single hypothesized TFBS for CTCF on a segment of the multiple alignment. For each sequence s the species from which it derives is shown on the right, and its WM score is shown on the left. The species with WM score larger than zero are selected (red sequences), the subtree involving these species of the full phylogenetic tree (red subtree of the black tree on the right) is obtained, and the probability of the selected sequences is calculated under an evolutionary model that incorporates selective constraints set by the WM. A sequence logo [16] of the CTCF motif is shown below the alignment.

models that incorporate rates of gain and loss of sites along different branches of the tree, we found that the most robust results are obtained using the following *species selection* procedure.

We follow the generally made assumption that TFs bind DNA in a fixed configuration, so that TFBSs for a single TF have a fixed length. Consequently, only species that are gaplessly aligned with respect to the reference can have an orthologous TFBS at the same location in the alignment. For example, in Fig. 3.2 the alignment implies that no orthologous site appears in dog. For every species that is gaplessly aligned relative to the reference, MotEvo calculates the probability of its sequence s under the WM and under a background model b. The probability $P(s|w)$ of a sequence segment s under the WM w is simply given by the product of WM-components, i.e.

$$P(s|w) = \prod_{i=1}^{l} w^i_{s_i}, \tag{3.1}$$

where l is the length of the WM and s_i is the nucleotide occurring at position i in sequence s.

MotEvo allows for a variety of background models. In the simplest model, there are 4 parameters b_α representing the probabilities for a nucleotide α to occur at a background position. MotEvo also allows k-th order background models in which the probability of a nucleotide depends on the k preceding nucleotides, i.e. 4^{k+1} conditional probabilities $P(s_i|s_{i-1}s_{i-2}\dots s_{i-k})$ that are estimated from the input sequences by default. For the simple single-nucleotide background model $P(s|b)$ is given by replacing the WM entries $w^i_{s_i}$ with the corresponding background probabilities b_{s_i} in equation (3.1), and analogously for the higher order models.

We refer to the log-ratio $\log[P(s|w)/P(s|b)]$ as the *WM score* of sequence s. For the species selection procedure MotEvo selects all species for which the WM score is larger than zero (the red sequences in Fig. 3.2). The key assumption that MotEvo now makes is that, whatever the reason is for the apparent loss of

the TFBS from certain species, they should not contribute to the evolutionary evidence for a TFBS to occur at this position in the reference species. Specifically, we obtain the subtree T' that is defined by the subset of 'red' species (indicated in Fig. 3.2) and *replace* the probability ratio $P(S|w,T)/P(S|b,T)$ by the one obtained using only this subtree, i.e. by $P(S|w,T')/P(S|b,T')$. This ensures that the 'black' sequences in Fig. 3.2 do not contribute to the ratio $P(S|w,T)/P(S|b,T)$.

The probability ratio $P(S|w,T)/P(S|b,T)$ is the product of independent contributions from the individual alignment columns, i.e.

$$\frac{P(S|w,T)}{P(S|b,T)} = \prod_{i=1}^{l} \frac{P(S_i|w^i,T)}{P(S_i|b,T)}, \tag{3.2}$$

where w^i denotes the ith WM column.

Imagine an alignment column that is evolving under the constraints set by WM column w^i and consider one branch of the phylogenetic tree T. Distances d along the branches of T are measured by the number of expected substitutions per neutrally evolving site. The key evolutionary quantities are the probabilities $P(\alpha|\beta,w^i,d)$ that, when evolving under WM column w^i along a branch of length d, a base β evolves into a base α. MotEvo uses a F81 model [17]. In this model the transition probabilities are given by

$$P(\alpha|\beta,w^i,d) = \delta_{\alpha\beta}e^{-d} + w^i_\alpha\left(1-e^{-d}\right). \tag{3.3}$$

Although it is straight-forward to implement more sophisticated evolutionary models, such as the model of [18], in practice the F81 model behaves very similarly. We chose the F81 model for consistency with the UFE model calculations which require the F81 model to be computationally tractable (see below).

The probability $P(S_i|w^i,T)$ is given by multiplying the probabilities $P(\alpha|\beta,w^i,T)$ for the transitions at each of the branches of the tree, setting α to the corresponding nucleotide for branches leading to the leafs of the tree, and summing over the unknown nucleotides at all internal nodes of the tree. To calculate the sum over the nucleotides at the internal nodes we use recursion relations introduced by [17] (see supplementary materials for details).

For the single nucleotide background model the probability $P(S|b,T)$ is calculated entirely analogously, i.e. simply replacing the WM column w^i with the column of background frequencies b in the above equations. One novel feature of MotEvo is that it allows the use of higher order background models in a phylogenetic setting by estimating the sequence-context at internal nodes by averaging over their descendants in the tree (see supplementary materials for details).

3.2.3 Unidentified Functional Elements

Even though the number of TFs for which WM models are available is increasing steadily, the sequence-specificity of the large majority of TFs is still unknown for most model organisms. Consequently, within the input alignments, there are likely many binding sites for TFs that are not represented by the WMs in our set $\{w\}$. Moreover, these TFBSs will often show significant evolutionary conservation, i.e. much more than can be expected under the background model and this can have undesirable consequences, as illustrated in Fig. 3.3.

In this example a TFBS for the SRF motif is predicted with high probability, despite the fact that the sequences show very poor matches to the WM (with the exception of mouse). The reason for this pathological behavior is that algorithms that do not explicitly take into account that unknown functional elements may occur in the input sequences, are forced to choose at each position of the alignment between assuming that the alignment segment contains a TFBS for one of the WMs in the input set, or that the alignment segment contains neutrally evolving sequences. Given a segment that is much more conserved than can be expected under neutral evolution, such algorithms may thus assign a high posterior to a site for WM w occurring, even when the sequences in the segment poorly match the WM.

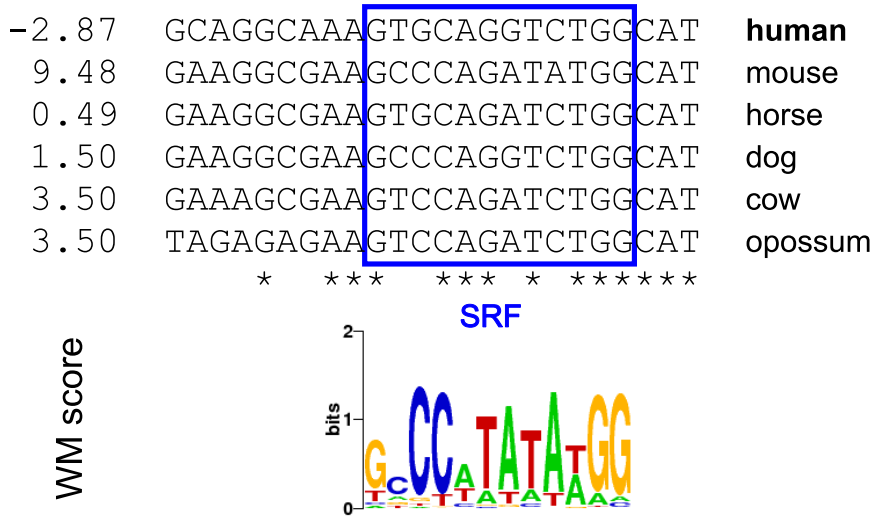


Figure 3.3: A small segment of a multiple alignment of orthologous mammalian sequences. The stars at the bottom of the alignment indicate columns that are perfectly conserved across all 6 species. A hypothetical binding site for the SRF motif is indicated (blue box) and the WM scores are shown for each of the sequences s. Without the UFE model, MotEvo assigns a posterior probability of 0.97 for a site for SRF to occur at this position, whereas with the UFE this probability drops to 0.01.

To avoid such spurious predictions MotEvo explicitly takes into account that the input alignments will contain well-conserved segments for motifs other than those in our input set $\{w\}$ which we call *Unidentified Functional Elements* (UFEs). To calculate the probability $P_{\text{ufe}}(S_k|T)$ of a single alignment column under the UFE we integrate the probability $P(S_k|w^k, T)$ over all possible WM columns w^k, i.e.

$$P_{\text{ufe}}(S_k|T) = \int P(S_k|w^k, T)P(w^k)dw^k, \tag{3.4}$$

where $P(w^k)$ is a prior distribution over possible alignment columns for which MotEvo uses a Dirichlet prior (see supplementary methods). To be able to calculate such integrals analytically, MotEvo uses the F81 model for the evolution along each branch of the tree. The supplementary materials provide details of the calculation of this integral.

Another parameter used by MotEvo is the length l_u of the UFE model, which is generally set to the typical length of TFBSs. The UFE model is then treated as any other WM, and the probability ratio $P_{\text{ufe}}(S|T)/P(S|b, T)$ is calculated for the l_u consecutive alignment columns of an hypothesized site of the UFE.

For the example shown in Fig. 3.3, when the UFE model of length 10 is used, the posterior of the SRF motif drops to 0.01 and sites for the UFE are predicted in this area with moderate posteriors. Note that MotEvo can also be run using *only* the UFE model. In this way, a conservation profile that quantifies the evidence for purifying selection across the alignments can be obtained without the need of providing specific motifs [19], i.e. providing a functionality similar to algorithms such as phastcons [20].

3.2.4 Forward/backward algorithm

In contrast to MONKEY [4] and other algorithms that scan with a single WM at a time, MotEvo predicts TFBSs for an arbitrary number of WMs and considers all possible configurations of non-overlapping TFBSs. Above we calculated the WM/background probability ratio $P(S|w, T)/P(S|b, T)$ for alignment segment S assuming a single hypothesized site for w. The probability ratio for an entire alignment given a configuration

containing multiple binding sites is simply the *product* of the ratios for each of the binding sites. Note that all parts of the alignment where no binding sites occur do not contribute to this ratio, i.e. their contributions cancel between numerator and denominator.

To assign prior probabilities to configurations, MotEvo assumes that scanning the reference species sequence from left to right, there is at each position a probability π_w for a site for WM w to start. For notational simplicity, we consider both the UFE model and the background model b as members of our set $\{w\}$ of WMs. The prior probability for a binding site configuration in which there are n_w sites for WM w is then proportional to

$$\prod_w (\pi_w)^{n_w} . \tag{3.5}$$

Note that we have the normalization condition $\sum_w \pi_w = 1$.

To calculate posterior probabilities we will need to sum over the probability ratios of all possible binding site configurations, i.e. calculate a partition sum. To this end we use recursion relations similar to those of the forward/backward algorithm used in the theory of hidden Markov models [21]. Let the sum of the probability ratios of all possible configurations of TFBSs up to position n in the reference species be denoted by F_n. Noting that any configuration ending at position n in the reference species has to end with a site for one of our WMs (which now include the background and UFE models) we have the recursion relation

$$F_n = \sum_{w \in \{w\}} \pi_w \frac{P(S_{[n-l_w,l_w]}|w,T)}{P(S_{[n-l_w,l_w]}|b,T)} F_{n-l_w}, \tag{3.6}$$

where l_w is the length of WM w.

Instead of moving from left to right over the multiple alignment, we can also move from right to left and define R_n as the sum over the probability ratios of all possible binding site configurations from position n until the end of the alignment. Further details are in the supplementary materials.

3.2.5 Transcription Factor Binding Site Predictions

Once the forward and backward sums F_n and R_n have been obtained we can calculate the posterior probabilities $P(w, n|S, \{w\}, T)$ that a binding site for WM w occurs at positions $n + 1$ through $n + l_w$:

$$P(w, n|S, \{w\}, T) = \frac{F_n \frac{P(S_{[n,l_w]}|w,T)}{P(S_{[n,l_w]}|b,T)} \pi_w R_{n+l_w+1}}{F_L}, \tag{3.7}$$

where F_L is the sum over probability ratios for all configurations for the entire alignment of length L. Note that the sum over all configurations in which a site for w occurs at n is equal to the sum over all possible configurations up to position n and all configurations from $(n+l_w+1)$ onwards. Using the procedures described above, posterior probabilities $P(w, n|S, \{w\}, T)$ at every position n for every WM w can be calculated in a time that is proportional to product of the length of the multiple alignment L, the number of WMs, and the number of species in the alignment. This linear scaling of the run-time allows MotEvo to make comprehensive site predictions for very large sequences using a large number of WMs in relatively short computational times. For example, when a TF is known to have different types of binding sites, e.g. half-sites separated by spacers of different lengths, MotEvo can easily run with WMs for each site type in parallel.

Note that, when running on a single sequence, MotEvo is equivalent to a statistical mechanical model that calculates the binding frequencies along the genome of multiple factors competing for binding along the genome, i.e. on a single sequence MotEvo is equivalent to the approach presented in [14]. In this biophysical interpretation the logarithms of the WM probabilities correspond to binding energies, the WM priors correspond to the concentrations of the different factors, and the posterior probabilities correspond to the fractions of time a given TF is bound at a given site.

3.2.6 Prior Updating

The prior probability distribution over binding site configurations is parametrized by the vector π that gives the expected binding site density π_w for each WM w, including the background and UFE models. This prior π can be specified by the user but MotEvo can also use an expectation-maximization algorithm to find the vector π that maximizes the probability of the observed alignments $\{S\}$.

We start with an initial prior vector π and calculate the posterior probabilities $P(w, n|S, \{w\}, T)$ for all alignments in the set $\{S\}$. We then calculate, for each WM, the sum n_w of the posterior probabilities $P(w, n|S, \{w\}, T)$ over all positions n in all alignments S. That is, n_w represents to the total expected number of binding sites for WM w. Using these MotEvo calculates a new prior vector

$$\pi_w = \frac{n_w}{\sum_{w' \in \{w\}} n_{w'}}, \tag{3.8}$$

and calculates new posterior probabilities $P(w, n|S, \{w\}, T)$ using this new prior vector. This procedure is iterated until the prior vector converges. It is easy to show, see e.g. [22], that this expectation-maximization procedure maximizes the probability of the input alignments with respect to the prior vector π.

3.2.7 Enhancer prediction

Enhancers are cis-regulatory elements on the genome that are distal to the promoter of the gene whose expression they regulate. They are characterized by a high density of TFBSs for a particular subset of TFs and they are typically a few hundred base pairs in length. They can occur both upstream, downstream, or in an intron of their target gene [23].

To find enhancers, MotEvo extends previously developed algorithms [12, 15] to multiple alignments with an arbitrary number of species and phylogenetic relationships. A window of a given length (typically a few hundred base pairs) is slid over the input alignments and for each window MotEvo predicts posterior probabilities of binding site occurrence for the set of input WMs. Importantly, MotEvo then updates the priors π_w *separately* for each window, allowing it to adapt the binding site densities to each window. Note that, because TFBSs cannot overlap within a single configuration, the prior updating roughly tries to maximize the number of TFBSs for the input WMs that can be bound at the same time to a given region. To assign a final *enhancer score* to a window, MotEvo calculates the log-ratio of the sum of the probabilities of all possible binding site configurations and the probability of the configuration with only background columns.

3.2.8 Weight Matrix Refinement

MotEvo also implements an expectation-maximization procedure for refining WMs based on its TFBS predictions. Formally, the idea is to maximize the probability of the entire input data $\{S\}$ with respect to the WMs $\{w\}$, starting from the WMs that were provided as input. As shown in the supplementary materials, this maximization can be obtained to a good approximation, using the following procedure.

Starting from the input WMs $\{w\}$ MotEvo first predicts posterior probabilities $P(w, n|S, \{w\}, T)$ for TFBS occurrence of each WM w at each position n in each input alignment S. It then calculates, for each WM w, each position i in the WM, and each nucleotide α, the sum $n^i_\alpha(w)$ of the posterior probabilities of all putative TFBSs that have a nucleotide α occuring at position i of the site in the reference species. That is, the statistics $n^i_\alpha(w)$ calculate the expected total number of TFBSs for WM w that, in the reference species, have nucleotide α at position i. Note that MotEvo can also be instructed to ignore sites with posteriors below a cut-off in calculating these sums. MotEvo then updates the WM as follows

$$w^i_\alpha = \frac{n_{i\alpha}(w)}{\sum_\beta n_{i\beta}(w)}. \tag{3.9}$$

Site predictions are then performed with these updated WMs and this procedure is repeated until the WMs converge. Using this procedure MotEvo thus extends the functionality of the PhyME algorithm [9], allowing for the simultaneous expectation-maximization of multiple motifs in parallel.

3.3 Results

Although a key benefit of MotEvo as a computational tool is that it integrates cutting-edge methods within a single executable, we here focus on evaluating the performance benefits of the novel features that MotEvo implements. For benchmarking, we collected 5 data-sets [24, 25] in which chromatin immuno-precipitation followed by next-generation sequencing (ChIP-seq) was performed for the human TFs CTCF, GABP, NRSF, SRF, and STAT1. Using the peak finder MACS [26] we determined the regions bound by the TF in question for each ChIP-seq data-set. Binding regions regions that occurred in more than one data-set were removed to ensure that (at least in the conditions tested) only one TF is bound to each region. Finally, we selected the 900 regions with highest enrichment from each data-set (see supplementary materials for details).

Using pairwise genome alignments from the UCSC data-base [27], we extracted orthologous regions from 6 other mammals (mouse, dog, cow, monkey, horse, and opossum) and obtained 7-way multiple alignments using T-coffee [28]. For NRSF, GABP, SRF, and STAT1 known WM motifs were taken from the literature [29] and a CTCF WM was inferred from ChIP-chip data in fly [30].

To test the new features we predict TFBSs with MotEvo on the benchmarking data-set both using the feature and with the feature turned off. We then compare the TFBS predictions and evaluate to what extent the TFBS predictions are able to infer which region was bound by which TF. In addition, we compared MotEvo's performance on these data-sets with those of MONKEY [31] and PhyloScan [32, 33]. Finally, we test the performance of motifs obtained using MotEvo's motif refinement and compare it with the performance of motifs inferred by MEME [8].

3.3.1 The UFE model strongly reduces spurious predictions

We argued above that, without the UFE, highly conserved regions are often mistakenly predicted as TFBSs for WMs in our set, even when the sequences poorly match the corresponding motif. To test this, we predicted TFBSs on all regions, once including the UFE and once without it. For each TF we determined the WM score of the sequence occurring in the reference species at each predicted TFBS, and constructed histograms of the distributions of WM scores (Fig. 3.7 and supplementary Fig. S1).

Inclusion of the UFE in general leads to substantial changes in the predicted TFBSs. First of all, the total number of predicted TFBSs is much lower with the UFE. Second, the UFE specifically causes predicted TFBSs that have a weak match to the motif to disappear. Finally, using the UFE the fraction of predicted TFBSs that fall in 'correct' regions, i.e. regions that were immuno-precipitated with the corresponding TF often increases dramatically, i.e. from around 40% to over 90% for 3 of the 5 TFs (insets of Fig. 3.7 and Fig. S1). However, from this test it is not clear if this increased specificity comes with a cost in sensitivity of the site predictions, which we address in the following test.

3.3.2 MotEvo's novel features improve TFBS prediction

To test MotEvo's performance, and the role of its novel features in a realistic setting, we tested how accurately the TFBS predictions can distinguish which region was bound by which of the 5 TFs. For each motif w and each region r we assign a score $n(r, w)$ by summing the posterior probabilities of all predicted TFBSs for w in r. We then obtain a sensitivity/positive-predictive value (PPV) curve by, as a function of a cut-off on the score $n(r, w)$, calculating the fraction of all regions bound by the corresponding TF that have a score above the cut-off (sensitivity) and the fraction of all regions with score above the cut-off that were indeed bound by the TF (PPV). We gather such sensitivity/PPV curves using MotEvo in its standard form, with the UFE model

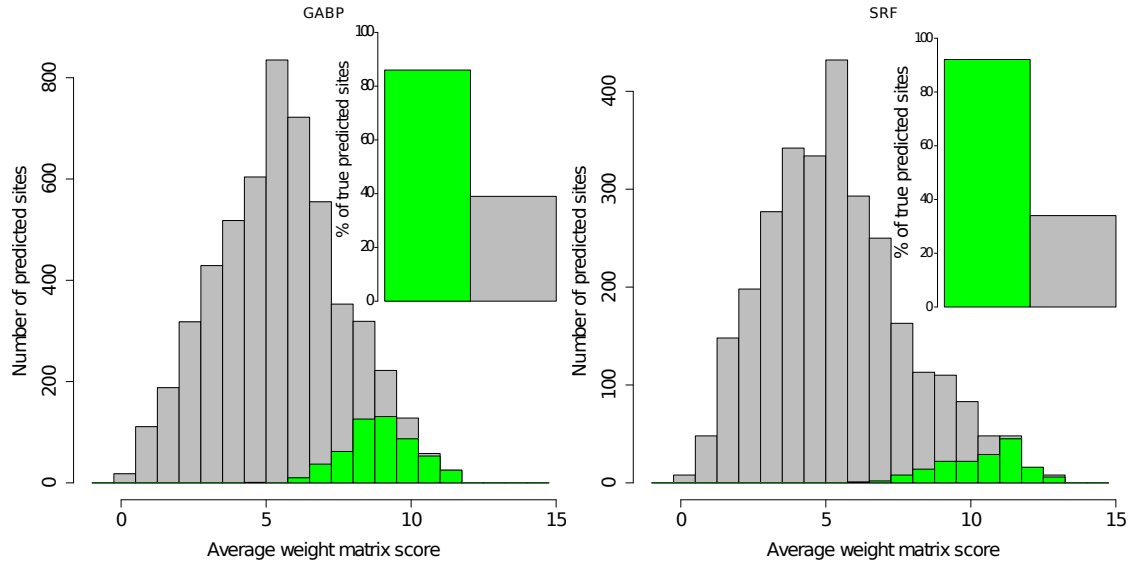


Figure 3.4: Comparison of TFBS predictions with and without inclusion of the UFE model. The histograms show the distribution of WM-scores (log-ratio $\log[P(s|w)/P(s|b)]$ for the sequence s occurring in the reference species) of the TFBSs predicted by MotEvo with (green) and without (grey) usage of the UFE. The insets show the fractions of predicted sites that fall within the 'correct' regions, i.e that were immuno-precipitated with the corresponding TF. Results are shown for the TFs GABP and SRF. Results for all TFs are shown in supplementary Figs. S1 and S2.

hturned off, and with species selection turned off, i.e. including sequences from all gaplessly aligned species at each putative site. We also obtained binding site predictions for two TFBS prediction algorithms that also incorporate an explicit evolutionary model: MONKEY [31] and PhyloScan [32, 33] (see supplementary methods for details), and determined their sensitivity/PPV curves.

We first of all see that, without the UFE, MotEvo's performance is dramatically reduced. In particular, because of the large number of spuriously predicted sites, no high specificity can be obtained without the UFE (Fig. 3.8 and supplementary Fig. S2). Only at very high sensitivities, i.e. when detecting even the regions with the weakest TFBSs, is the performance relatively unaffected. Besides the UFE, MotEvo explicitly considers that TFBSs may have been lost in a subset of the species, either through evolution or simply because of errors in the multiple alignment, by using the 'species selection' scheme described above. We find that species selection leads to an increase in performance for all TFs (supplementary Fig. S2) ranging from small improvements at some TFs (NSRF, CTCF) to moderate or even very large improvements for others (GABP and SRF, Fig. 3.8). Thus, MotEvo's method for treating loss and gain of TFBSs within the phylogeny significantly outperforms methods that assume that all sequences in the multiple alignment are evolving under the same selective constraints determined by the WM.

Although we invested some efforts optimizing the performance of MONKEY and PhyloScan on this data-set (see supplementary methods) MotEvo signifcantly outperforms these algorithms (Fig. 3.8 and supplementary Fig. S2). Especially striking is the inability of these algorithms to reach high sensitivity for motifs with high information content (NRSF, CTCF). Manual inspection of the differences in the predictions of MotEvo, PhyloScan, and MONKEY strongly suggest that MONKEY and PhyloScan's performance is most affected by their inability to deal with alignment segments where a binding site occurs in only some of the species, i.e. where some of the species have either gaps relative to the reference species or low WM scores (see supplementary materials for more discussion). These algorithms effectively assume that a functional site must appear in all species of the alignment, and for multiple alignments involving 7 species there are many cases where this is too restrictive an assumption.

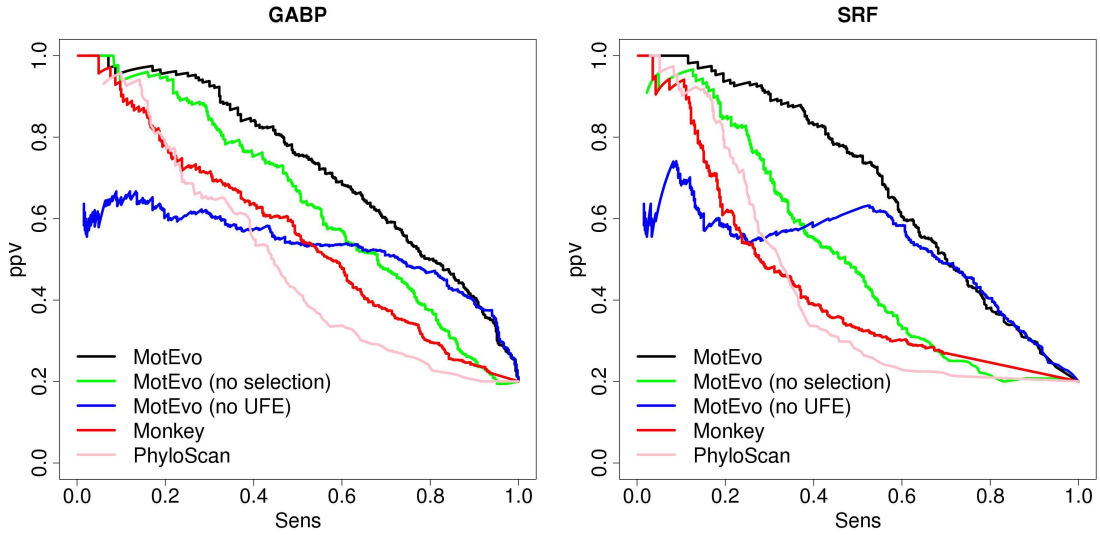


Figure 3.5: Comparison of TFBS prediction accuracy for MotEvo (black), MotEvo without the UFE (blue), MotEvo without species selection (green), Monkey (red), and PhyloScan (pink). TFBS predictions were made on the benchmark set of 5 times 900 regions by each of the methods and were then used to predict, for each TF, which of the 4500 regions were bound by the TF. The panels show sensitivity/positive-predictive value curves for the performance obtained predicting the regions bound by the TFs STAT1 and GABP. Results for all TFs are shown in supplementary Fig. S2.

Finally, we checked whether MotEvo's performance is strongly affected by the alignment algorithm used (supplementary Fig. S6) and find that sensitivity/PPV curves change only marginally when using different alignment methods. Similarly, use of a higher-order background model also only marginally improves performance on these benchmarking data-sets (supplementary Fig. S6).

3.3.3 WM refinement improves TFBS predictions

Through the availability of next-generation sequencing technologies many labs have started performing ChIP-seq of TFs of interest, and the number of available data-sets is increasing rapidly. As ChIP-seq is able to identify large numbers, i.e. hundreds to thousands, of binding regions for a given TF genome-wide, this offers the possibility to investigate the binding specificity of TFs at much higher levels of resolution than was previously possible.

MotEvo implements an expectation-maximization strategy for refining WMs from an input data-set of binding regions which we here test using the same benchmarking ChIP-seq data for 5 TFs. For each of the 5 TFs we randomly selected 450 of the 900 peaks and pooled them into one large data-set that we used as a *test set*. The remaining 450 regions for each TF were used as a *training set* for WM refinement. Besides MotEvo, we also used MEME [8] to infer a WM motif for each of the 5 test-sets. Sequence logos of the original and inferred motifs are shown in Fig. 3.9 and supplementary Figs. S3, and S4.

To test the performance of these WMs in identifying which TF binds to which target region, we predicted TFBSs for all original and refined WMs on the *test-set* using MotEvo. As in the previous section, we assign a score $n(r, w)$ for each WM w to each region r by summing the posterior probabilities of predicted TFBSs, and obtain sensitivity/PPV curves that quantify the performance of the TFBSs in predicting which TF was bound by each region (Fig. 3.9 and supplementary Fig. S3).

The improvement that refinement provides over the literature motif ranges from virtually no difference between original and refined WMs (GABP), through moderate improvements (SRF), to very dramatic im-

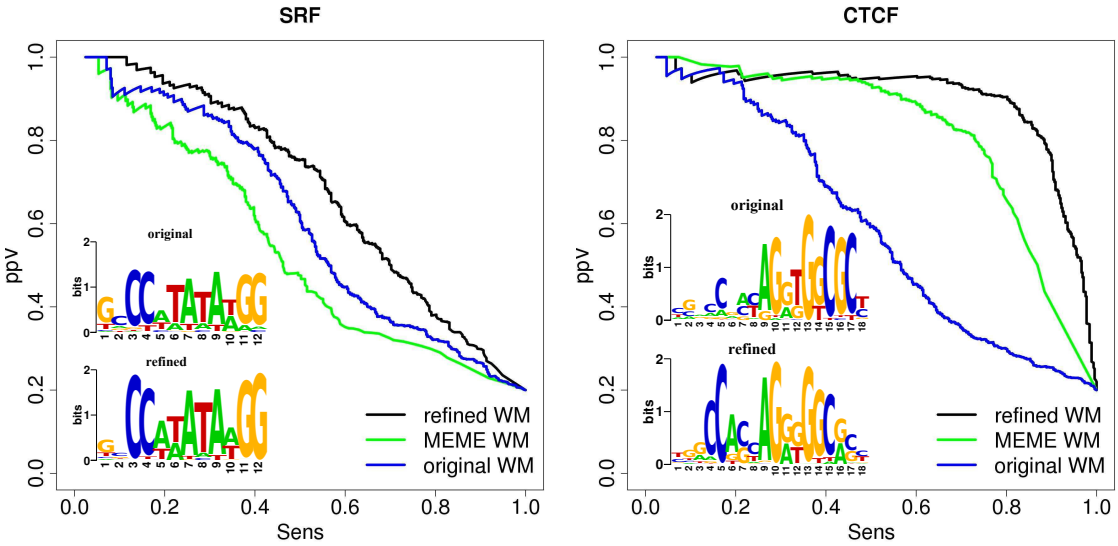


Figure 3.6: Comparison of the performance in predicting TF binding of original WMs based on literature, WMs refined by MotEvo, and WMs inferred by MEME. Binding sites were predicted by MotEvo on all 5×450 regions in the test-set using the three WMs for each of the 5 TFs. The predicted TFBSs were used to predict which TF is bound by each region. The panels show sensitivity/positive-predictive value curves for the performance obtained using the original (blue), the refined motif (black), and MEME's motif (green) for the TFs SRF and CTCF. Sequence logos of the original and refined WMs are shown in each panel as well. Results for all TFs are shown in supplementary Fig. S3.

provements (CTCF). It is notable that, even when there is a clear difference in performance of the original and refined WMs, the sequence logos appear very similar visually. This illustrates that subtle changes in the WM can have substantial effects on TFBS predictions. Interestingly, in the case of STAT1 the refined motif is closer to a true palindrome than the original motif, suggesting that these sites are bound by the TF in dimer form. This is supported by the literature, i.e. it is known that STAT1 is 'activated' through phosphorylation at a tyrosine, after which STAT1 proteins form homo-dimers that are translocated to the nucleus to regulate transcription [34]. The most dramatic improvement in performance is observed for CTCF. This is not surprising given that, in contrast to the other TFs in this set, for CTCF the original WM was based on data from Drosophila. That is, it is plausible that the precise sequence specificity of CTCF may differ between Drosophila and human.

MotEvo's refined WMs outperform the WMs inferred by MEME for all TFs (Fig. 3.9 and supplementary Fig. S3). Although the difference is marginal for NRSF, for some of the TFs MEME's motif performs clearly worse than the literature motif (see supplementary materials for further discussion). In summary, our results show that MotEvo's WM refinement consistently improves the ability of the WM to distinguish regions bound by the TF from regions that are bound by other TFs. In addition, this improvement can be very large in some cases.

3.3.4 Enhancer prediction accuracy increases with the number of species used

Finally, we evaluated MotEvo's ability to predict distal cis-regulatory modules (also called enhancers). For testing we used the set of 76 experimentally validated blastoderm cis-regulatory modules (CRMs) that was collected in [35]. The length of these CRMs ranges from around 90bps to 2kbps. For each CRM we extracted the flanking region of 5kb before and after the CRM on the genome. We then multiply-align these regions with other sequenced Drosophila species. As it is likely that these regions contain other blastoderm CRMs

Table 3.1: Overlap between the known and predicted blastoderm cis-regulatory modules as a function of the number of Drosophila species used.

Number of species	1	2	3	5	7	9
Performance	0.57	0.70	0.73	0.76	0.82	0.93

that are unknown, we shuffle the aligned columns of the flanking region without changing the conservation pattern, that is to say we substitute each column by a column with similar conservation (same gap pattern and subset of species that has the same base as the reference). Exonic regions are also excluded.

To test how prediction accuracy depends on the number of species used in the multiple alignments we made several pruned versions of the multiple alignments by selecting different subsets of the available species, ranging from only the melanogaster sequence, to sequences from 9 available species (see supplementary materials and Fig. S5 for details). We use MotEvo with 7 WMs for Drosophila TFs known to be involved in binding to these enhancers (Bcd, Cad, Dl, Hb, Kni, Kr, and Tll), to perform enhancer predictions on all multiple alignments, selecting for each alignment the $900bp$ window with the highest enhancer score as the predicted enhancer. To assess the performance we calculated, for each alignment, the overlap between the predicted and the known enhancer. As table 3.1 shows, the performance increases significantly as the number of species increases, reaching over 90% performance when 9 species are used. This illustrates that MotEvo's ability to predict enhancers on multiple alignments of an arbitrary number of species significantly improves accuracy over methods that only allow pairwise analysis.

3.4 Discussion

From a theoretical point of view, the major advantage of the MotEvo method we presented here is that it integrates cutting-edge Bayesian probabilistic methods for the prediction of TFBSs, regulatory motifs, and conservation patterns within one consistent theoretical frame work, developing several novel features such as the UFE model and species selection in the process. In addition, our benchmarking tests have demonstrated that these features improve MotEvo's performance, and that MotEvo outperforms currently available methods. Another major advantage of the MotEvo tool is its versatility, i.e. by simple changes to the parameter file the tool can perform a wide array of tasks ranging from motif inference, to enhancer prediction, to conservation profile mapping, site density estimation, and of course TFBS prediction. Moreover, essentially all variables used by the algorithm, from phylogeny to background models, to priors, can be controlled by the user, allowing these to be adapted to a wide range of applications. For example, using MotEvo in combination with genome-wide mapping of transcription start sites we have predicted functional TFBSs for hundreds of WMs in proximal promoters in human, mouse [36], yeast [37], and E. coli. We have also obtained refined WMs using MotEvo on ChIP-seq data-sets for a number of TFs beyond those studied here. All these TFBS and motif predictions are available for download from our SwissRegulon database at www.swissregulon.unibas.ch.

As experimental validation of the functionality of individual TFBSs is extremely labor intensive, it remains highly challenging to estimate the accuracy of large-scale TFBS predictions. In the past it has sometimes been assumed that ChIP-chip and ChIP-seq data-sets can be treated as a gold standard, but our own analysis suggests that computational predictions can in fact be considerable more accurate in mapping functional sites than such high-throughput experimental approaches [37].

What is clear is that TF binding and function is highly context dependent. For a TF with a short degenerate motif there may be millions of sites genome-wide with motif matches at least as high as known functional sites, but in a typical ChIP-seq experiment only a few thousand of these are found to be actually bound, and even among these only a small subset may directly affect gene expression. In the search for variables that provide important context it is important to distinguish those that are merely *predictive* for the functionality of TFBSs from those that are *explanatory*. Cross-species conservation is an example of an explanatory variable,

i.e. highly conserved TFBSs are more likely to be functional, but conservation is not explanatory; sequences in other species cannot explain why a particular sequence is bound or functional in a given species.

In higher eukaryotes the chromatin state is likely to be an important explanatory variable. In areas where the nucleosomes are densely packaging the DNA, it may be hard for a TF to access an individual site in the DNA. This may explain why functional TFBSs come in clusters of nearby sites for co-expressed TFs: these TFs may passively cooperate in displacing the nucleosomes from the DNA. Such a model can potentially explain the observation that the genomic binding pattern observed for a given TF is dependent on the expression profiles of other TFs [38]. However, to what extent binding and function of TFBSs is dependent on a high-order 'grammar' of TFBS configurations, i.e. the precise spacing and relative orientation of the sites, is currently unclear. In order to make progress on these important questions we believe that the largest potential lies in integrating sequence analysis with the analysis of temporal patterns of TF binding, of genome-wide chromatin states, and of the expression of potential target genes.

Acknowledgement

This work was supported by Swiss National Science Foundation grant 31003A-118318 and SystemsX.ch through the CellPlasticity project.

3.5 Supplementary Methods

In this section we discuss several of the methods used by MotEvo in more detail.

3.5.1 Likelihood of an alignment column

Here we explain how we use the Felsenstein recursion relations [17]. An alignment segment S is hypothesized to contain a binding site for WM w and we want to calculate the probability $P(S_i|w^i, T)$ of the alignment column S_i, i.e. the ith column of the alignment, given WM column w^i and phylogenetic tree T. As described in the main article, we assume that, along each branch of the tree, the probability to evolve from ancestor nucleotide α to descendent nucleotide β is given by

$$P(\alpha|\beta, w^i, d) = \delta_{\alpha\beta} e^{-d} + w^i_\alpha \left(1 - e^{-d}\right), \tag{3.10}$$

where d is the length of the branch, and $\delta_{\alpha\beta}$ is the Kronecker delta function that is 1 when $\alpha = \beta$ and 0 otherwise. The probability of the column S_i for the entire phylogenetic tree can now be calculated recursively as follows.

Let n denote a node of the phylogenetic tree and let $S_i(n)$ denote the nucleotides of alignment column S_i that stem from species that are descendants of node n in the tree. We let $P(S_i(n)|\alpha, w^i, T)$ denote the probability of observing nucleotides $S_i(n)$ assuming that node n had nucleotide α. First, for a node n that is a leaf of the tree, we have that α must match the nucleotide s_n that occurs at this leaf, i.e.

$$P(S_i(n)|\alpha, w^i, T) = \delta_{s_n\alpha}. \tag{3.11}$$

For an internal node of the tree we have the following recursion relation

$$P(S_i(n)|\alpha, w^i, T) = \prod_{m \in D(n)} \left[\sum_\beta P(\beta|\alpha, w^i, d_m) P(S_i(m)|\beta, w^i, T) \right], \tag{3.12}$$

where the product is over the set $D(n)$ of direct descendants of node n in the tree, and d_m is the distance from node m to its parent node n. What this equation says is that, the probability for all data below node n given that nucleotide α occurs at this node, is given by a product of probabilities for all data at each of the descendants m of node n. For each descendant, the probability is given by, summing over all possible nucleotides β at the descendant, the transition probability from α to β times the probability $P(S_i(m)|\beta, w^i, T)$ of the data below descendant m given nucleotide β. Starting from the leafs and moving up the tree we can use this recursion relation to efficiently calculate the probabilities $P(S_i(n)|\alpha, w^i, T)$ for each node n up to the root r of the tree. Finally, at the root we have

$$P(S|w^i, T) = \sum_\alpha P(S_i(r)|\alpha, w^i, T) w^i_\alpha, \tag{3.13}$$

where we have used that, under the assumption that the sequences are binding sites for the WM, the prior probability that the root sequence has letter α at position i is simply the WM entry w^i_α.

3.5.2 Higher order background models within the evolutionary model

To calculate the probability $P(S_i|b, T)$ for an alignment column S_i with a simple background model, i.e. given by the 4 nucleotide probabilities b_α, we can use precisely the same equations as we used for calculating $P(S_i|w^i, T)$. We simply replace the weight matrix entries w^i_α with the background probabilities b_α. In particular, the transition probabilities on each branch of the tree are given by

$$P(\alpha|\beta, b, d) = \delta_{\alpha\beta} e^{-d} + b_\alpha \left(1 - e^{-d}\right), \tag{3.14}$$

where d is again the length of the branch.

When using higher order background models, the background probabilities for individual sequences depend on the immediately preceding bases and these will typically differ across the species in the alignment. Therefore, we generally do not know what the flanking bases in the internal nodes were, nor when during the evolution the flanking bases underwent substitutions. In a rigorous treatment we would have to consider all possible scenarios for evolution of the flanking bases. However, the evolution of the flanking bases would again depend on the evolution of *their* flanking bases, and so on, effectively coupling together the evolution of all columns in the entire alignment. As a consequence, such a rigorous is not computationally tractable, as far as we are aware, and we use the following practical approximation.

In order to be able to apply the recursion relations (3.12) we essentially need, for each branch of the tree, a set of 4 background probabilities b_α to use in equation (3.14). For the branches leading to the leafs of the tree, we use the k-th order background model to construct a vector of 4 background frequencies b using the sequence context at the leafs. That is, we are effectively assuming that, for each branch leading to a leaf, the sequence-context was the same for the entire evolutionary branch leading up to their ancestor. For a branch leading to an internal node of the tree we similarly construct a background vector b by *averaging* the background frequency vectors b that were used for the branches leading to the direct descendants of the node. For example, for a branch that is the ancestor of 2 leaf nodes we simply average the background vectors b for these two leafs, i.e. as given by the sequence-contexts at the leafs. The background vectors b for each internal node can thus also be determined recursively, starting at the leafs of the tree and moving upward toward the root.

3.5.3 Calculating the probabilities under the UFE model

We have shown above how the probability for an alignment column S_k given a WM column w^k and phylogenetic tree T can be calculated using our F81 evolutionary model together with the Felsenstein recursion relations. To calculate the probability $P_{\mathrm{UFE}}(S_k|T)$ of alignment column S_k under the UFE model we integrate over all possible WMs:

$$P_{\mathrm{ufe}}(S_k|T) = \int P(S_k|w^k, T)P(w^k)dw^k, \tag{3.15}$$

where $P(w^k)$ is a prior distribution over possible alignment columns, and the integration is over the simplex of all possible vectors w^k such that all components are non-negative ($w^k_\alpha \geq 0$) and normalized ($\sum_\alpha w^k_\alpha = 1$).

Before we discuss calculating this integral, we first discuss the choice of prior. To keep the integrals computationally tractable we use a standard Dirichlet prior of the form:

$$P(w^k) \propto \prod_\alpha (w^k_\alpha)^{\lambda_\alpha - 1}, \tag{3.16}$$

where the λ_α are a set of four parameters that determine the shape of the prior. As we have shown previously [19], consistency in the limit of $d \to 0$ (short evolutionary branches) requires that the parameters λ_α are proportional to the background probabilities b_α, i.e. $\lambda_\alpha = \lambda b_\alpha$. The parameter λ can still be chosen and determines the information content that is expected for WM columns *a priori*. We use $\lambda = 1$, which roughly corresponds to the average information content per WM column for experimentally determined TFBSs in E. coli. That is, we use $\lambda_\alpha = b_\alpha$.

We now turn to the integral. Provided that we use the F81 model for the evolutionary transition probabilities $P(\alpha|\beta, w^k, d)$, the integral (3.15) can be performed analytically. As mentioned previously, if we assume the nucleotides at the internal nodes of the tree fixed, then the probability $P(S_k|w^k, T)$ given these nucleotides is simply a product over transition probabilities $P(\alpha|\beta, w^k, d)$ for each branch of the tree. Each of these transition probabilities $P(\alpha|\beta, w^k, d)$ is either directly proportional to w^k_α (when $\alpha \neq \beta$) or a sum

of a constant term and a term proportional to w_α^k (when $\alpha = \beta$). Thus, the product is a general *polynomial* in the WM entries w_α^k. Such a polynomial can be written as a sum over monomial terms, i.e

$$\sum_i c(i,T) \prod_\alpha \left(w_\alpha^k\right)^{n_\alpha(i)}, \tag{3.17}$$

where the indices i refer to the terms of the polynomial, $c(i,T)$ is a constant prefactor that depends on the branch lengths in the tree T, and the $n_\alpha(i)$ are integer exponents associated with term i. We now note that, since the product of transition probabilities takes on this general form given fixed nucleotides at the internal nodes, so does the general expression $P(S_k|w^k,T)$. That is, when we sum over all possible nucleotides at the internal nodes we just obtain a lot more terms in the above sum, but the general form of the expression remains unchanged. That is, we can write $P(S_k|w^k,T)$ as a general polynomial

$$P(S_k|w^k,T) = \sum_i c(i,T) \prod_\alpha \left(w_\alpha^k\right)^{n_\alpha(i)}. \tag{3.18}$$

We now make use of the fact that the integral of each monomial term has a simple solution, i.e.

$$\int \prod_\alpha \left(w_\alpha^k\right)^{n_\alpha(i)+\lambda b_\alpha - 1} dw^k = \frac{\prod_\alpha \Gamma\left[n_\alpha(i)+\lambda b_\alpha\right]}{\Gamma\left[\lambda + \sum_\alpha n_\alpha(i) + b_\alpha\right]}. \tag{3.19}$$

Thus, to determine $P_{\rm ufe}(S_k|T)$ we in principle just need to explicitly summing over all terms in the polynomial (3.18), i.e we obtain

$$P_{\rm ufe}(S_k|T) = \sum_i c(i,T) \frac{\prod_\alpha \Gamma\left[n_\alpha(i)+\lambda b_\alpha\right]}{\Gamma\left[\lambda + \sum_\alpha n_\alpha(i)\right]}. \tag{3.20}$$

Note, however, that the number of terms i increases exponentially with the number of species in the alignment. Although there is no simple recursion relation that reduces the calculation from exponential to polynomial time, we can use a variant of the Felsenstein recursion relations to efficiently calculate the pre-factors $c(i,T)$ for all polynomial terms i that have different counts in their exponents $n_\alpha(i)$.

Because the calculation of the UFE model probabilities $P_{\rm ufe}(S_k|T)$ is computationally expensive we typically proceed as follows. Given a phylogenetic tree T with N species, and a vector of background probabilities b_α, MotEvo can be asked to calculate the probabilities $P_{\rm ufe}(S_k|T)$ for all 4^N possible alignment columns S_k and save the results in a file. In other runs MotEvo can then be directed to use this file to read in the UFE probabilities rather than re-calculating them from scratch. That is, for each set of species with a given phylogenetic tree, we only have to calculate the UFE model probabilities once.

3.5.4 Backward recursion relation

We define R_n as the sum over the probability ratios of all possible binding site configurations from position n until the end of the alignment. These 'backward' partition sums can be determined using a recursion relation very similar to that for the forward partition sum

$$R_{n+1} = \sum_{w\in\{w\}} \pi_w \frac{P(S_{[n,l_w]}|w,T)}{P(S_{[n,l_w]}|b,T)} R_{n+l_w+1}. \tag{3.21}$$

3.5.5 Weight matrix refinement

We here derive the equations of the WM refinement procedure that MotEvo employs. We denote by $P(\{S\}|\{w\},T)$ the probability of the set of multiple alignments $\{S\}$ given the set of inputs WMs $\{w\}$ and the phylogenetic tree T. We want to maximize this probability $P(\{S\}|\{w\},T)$ with respect to all the WMs. To this end we

will have to calculate the derivatives of this probability with respect to the entries $\tilde{w}^k_\alpha$ of the kth WM column $\tilde{w}^k$ of a particular WM $\tilde{w}$ from the set $\{w\}$, for each column and for each WM.

First, we note that the probability $P(\{S\}|\{w\},T)$ is a product over the probabilities of the individual alignments S in the set, i.e.

$$P(\{S\}|\{w\},T) = \prod_{S\in\{S\}} P(S|\{w\},T). \tag{3.22}$$

Second, we can explicitly write these probabilities in terms of sum over all possible binding site configurations C

$$P(S|\{w\},T) = \sum_C P(C|\pi)P(S|C,\{w\},T), \tag{3.23}$$

where $P(C|\pi)$ is the prior probability assigned to configuration C, which depends on the WM priors π_w. If we denote the number of binding sites for WM w in configuration C by $n_w(C)$ this prior is given by

$$P(C|\pi) = \prod_w (\pi_w)^{n_w(C)}, \tag{3.24}$$

where we note that, for notational simplicity, the UFE and background model are considered members of the WM set $\{w\}$.

The probability $P(S|C,\{w\},T)$ of the alignment given the configuration C can itself be written as

$$P(S|C,\{w\},T) = \prod_w \prod_{i\in C_w} P(S_{[i,l_w]}|w,T), \tag{3.25}$$

where C_w denotes the set of positions in the multiple alignment where, in configuration C, a TFBS for WM w starts. Using this expression, we find for the derivative of $P(S|C,\{w\},T)$ with respect to a particular WM entry $\tilde{w}^k_\alpha$

$$\frac{\partial P(S|C,\{w\},T)}{\partial \tilde{w}^k_\alpha} = P(S|C,\{w\},T) \sum_{i\in C_{\tilde{w}}} \left[\frac{1}{P(S_{[i,l_{\tilde{w}}]}|\tilde{w},T)} \frac{\partial P(S_{[i,l_{\tilde{w}}]}|\tilde{w},T)}{\partial \tilde{w}^k_\alpha} \right]. \tag{3.26}$$

The derivative on the right in equation (3.26) gives the derivative of the probability particular alignment segments (hypothesized in configuration C to be a TFBS for WM $\tilde{w}$) with respect to a WM component. To understand the meaning of this derivative we consider what it gives in a few simple limits. First, let us assume that the alignment segments contains *only* the reference species, i.e. the segment equals the sequence s that occurs in the reference. Using $P(s|w) = \prod_i w^i_{s_i}$ we find

$$\frac{1}{P(s|\tilde{w},T)} \frac{\partial P(s|\tilde{w},T)}{\partial \tilde{w}^k_\alpha} = \frac{\delta_{s_k\alpha}}{w^k_\alpha}. \tag{3.27}$$

That is, the derivative is simply equal to one 1 over the WM entry when the kth nucleotide in the sequence s equals α and zero otherwise. Another simple limit that we may consider is that the alignment segment S contains sequences from N species that are sufficiently diverged such that they an be considered phylogenetically independent. In that limit we may ignore the tree T and approximate $P(S|w,T)$ by

$$P(S|w,T) = \prod_{i=1}^{l_w} \prod_\alpha (w^i_\alpha)^{n^i_\alpha(S)}, \tag{3.28}$$

where $n^i_\alpha(S)$ is the number of species whose sequence has nucleotide α occurring at position i. It is easy to see that in this limit we have simply

$$\frac{1}{P(S|\tilde{w},T)} \frac{\partial P(S|\tilde{w},T)}{\partial \tilde{w}^k_\alpha} = \frac{n^k_\alpha(S)}{w^k_\alpha}. \tag{3.29}$$

In summary, in these two limits the general form of this derivative is simply the number of sequences in S that have letter α at position k divided by the WM entry itself. We now make the assumption that this general form extends to the more complicated phylogenetic cases. Although this is strictly not true, we previously showed [5] that this approximation gives accurate results. We thus assume that, given the phylogenetic tree T and alignment S, there is an 'effective' number of species $n_\alpha^k(S,T)$ that have nucleotide α at position k, and assume that the derivative can be approximated by

$$\frac{1}{P(S|\tilde{w},T)}\frac{\partial P(S|\tilde{w},T)}{\partial \tilde{w}_\alpha^k} = \frac{n_\alpha^k(S,T)}{w_\alpha^k}. \tag{3.30}$$

We use this result to derive a more intuitive expression for the derivative of the logarithm of the total probability $P(\{S\}|\{w\},T)$ with respect to a single WM entry $\tilde{w}_\alpha^k$. We find

$$\frac{\partial P(\{S\}|\{w\},T)}{\partial \tilde{w}_\alpha^k} = \sum_{S\in\{S\}} \frac{1}{P(S|\{w\},T)} \sum_C P(S|C,\{w\},T) \sum_{i\in C_{\tilde{w}}} \frac{n_\alpha^k(S_{[i,l_{\tilde{w}}]},T)}{\tilde{w}_\alpha^k}. \tag{3.31}$$

We can further simplify by noting that, for each position i, the weighted sum over probabilities corresponds exactly to the *posterior* probability $P(i,\tilde{w}|S,\{w\},T)$ that a TFBS for WM $\tilde{w}$ occurs position i of alignment S. We then have

$$\frac{\partial P(\{S\}|\{w\},T)}{\partial \tilde{w}_\alpha^k} = \frac{1}{\tilde{w}_\alpha^k} \sum_{S\in\{S\}} \sum_i P(i,\tilde{w}|S,\{w\},T) n_\alpha^k(S_{[i,l_{\tilde{w}}]},T). \tag{3.32}$$

But if we think of $n_\alpha^k(S,T)$ as the number of sequences in segment S that have letter α at position k, then the above sum is simply the total number of expected occurrences of letter α in position k of predicted TFBSs for WM $\tilde{w}$.

Finally, to optimize Wm column $\tilde{w}^k$ under the normalization constraint $\sum_\alpha w_\alpha^k = 1$ we need to have that all derivatives equal the same constant, i.e.

$$\frac{\partial P(\{S\}|\{w\},T)}{\partial \tilde{w}_\alpha^k} = \text{constant} \tag{3.33}$$

Using this we obtain

$$\tilde{w}_\alpha^k = \text{constant} \sum_{S\in\{S\}} \sum_i P(i,\tilde{w}|S,\{w\},T) n_\alpha^k(S_{[i,l_{\tilde{w}}]},T). \tag{3.34}$$

That is, the WM entry should be proportional to a weighted sum over all putative TFBSs for WM $\tilde{w}$, weighing each site both by its posterior and by the number of independent occurrences of nucleotide α at position k occur within the corresponding alignment segment. As a final approximation, we know replace the function $n_\alpha^k(S_{[i,l_{\tilde{w}}]},T)$ with the one that would be obtained if only the reference sequence were present. That is, we set $n_\alpha^k(S_{[i,l_{\tilde{w}}]},T) = \delta_{\alpha s_k}$, where s_k is the nucleotide that occurs at position k in the sequence segment from the reference species. That is, although the full effect of all other species in the alignment is incorporated into the posterior probability of the site, we effectively ignore the other species in the second factor. Although we are losing information under this approximation, as long as there is no systematic bias in the TFBS sequences coming from different species, the final predictions should not be affected by this approximation.

3.6 Construction of benchmarking regions

For bench-marking, we collected 5 data-sets [24, 25] in which chromatin immuno-precipitation followed by next-generation sequencing (ChIP-seq) was performed for the human TFs CTCF, GABP, NRSF, SRF, and

STAT1. For all five data sets, the reads had already been filtered (quality controlling) and mapped to the genome. Using the peak finder MACS [26] we determined the regions that bound the TF in question for each ChIP-seq data-set. Binding regions were extended/reduced with respect to their middles such that they all had a length of 400bps. Binding regions that occurred in more than one data-set were removed to make sure that only one TF is bound to each region. Finally, we selected the 900 regions with highest enrichment according to MACS' reported p-value from each data-set. Using pairwise genome alignments from the UCSC database [27], we extracted orthologous regions from six other mammals (mouse, dog, cow, monkey, horse, and opossum) and obtained 7-way multiple alignments using T-coffee [28]. For NRSF, GABP, SRF, and STAT1 known WM motifs were taken from the literature [29] and a CTCF WM was inferred from ChIP-chip data in fly [30].

For the WM refinement evaluation, we randomly selected 450 (of the above mentioned) regions for a given TF and used it as a training set, that is to say those were the regions we refined the motif on. The remaining 450 regions were used as a test set to create sensitivity/positive-predictive value curves.

3.7 The UFE model strongly reduces spurious predictions

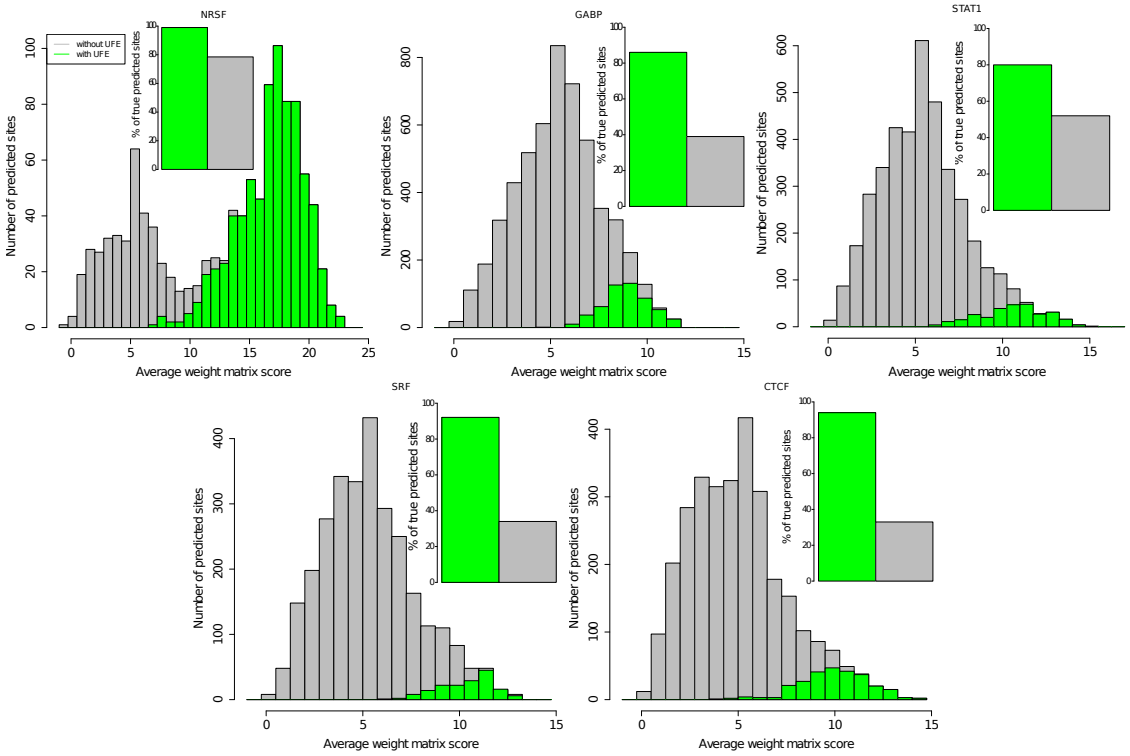


Figure 3.7: Comparison of TFBS predictions with and without inclusion of the UFE model. The histograms show the distribution of WM-scores (log-ratio $\log[P(s|w)/P(s|b)]$ for the sequence s occurring in the reference species) of the TFBSs predicted by MotEvo with (green) and without (grey) usage of the UFE. The insets show the fractions of predicted sites that fall within the 'correct' regions, i.e that were immuno-precipitated with the corresponding TF.

3.8 Species selection improves TFBS predictions

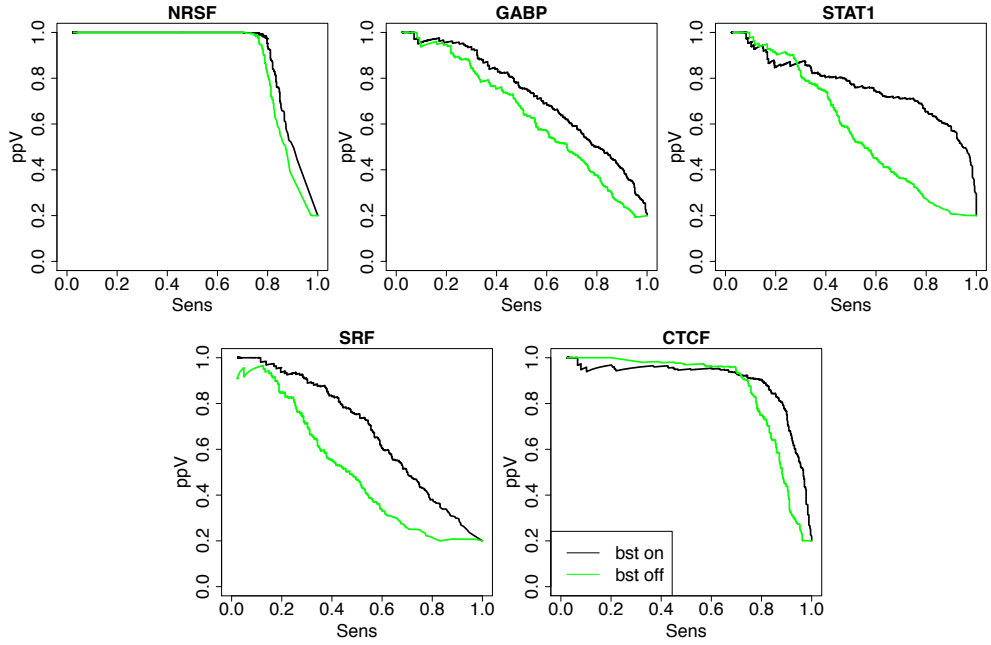


Figure 3.8: Species selection improves TFBS prediction accuracy. TFBS predictions were made on the bench-mark set of 5 times 900 regions by MotEvo once using species selection, i.e. selecting at each putative site only the species with sequences that score higher for the WM than for the background model, and once without it, i.e. including all gaplessly aligned species per definition. The TFBS predictions were used to predict, for each TF, which of the 4500 regions were bound by the TF and the panels show sensitivity/positive-predictive value curves for the performance obtained using species selection (black) and without it (green).

3.9 WM refinement

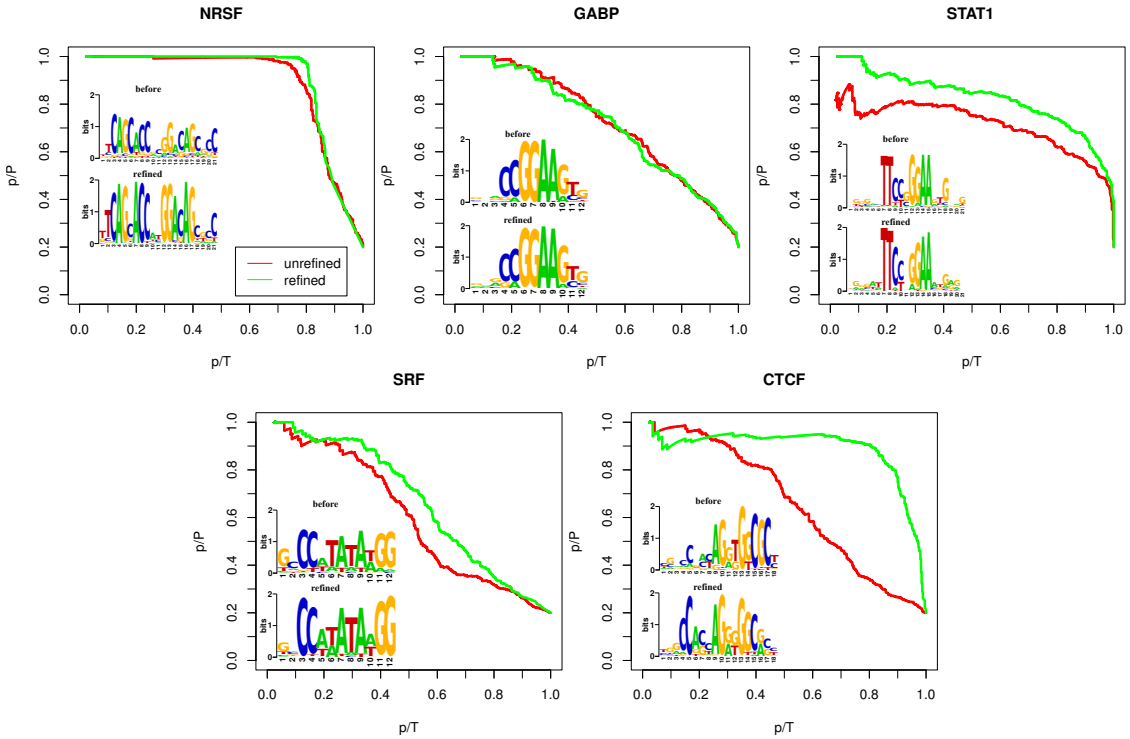


Figure 3.9: Comparison of the performance of original and refined WMs in predicting TF binding. Binding sites were predicted by MotEvo on all 5×450 regions in the test-set using both the original and refined WMs for each of the 5 TFs. The predicted TFBS were used to predict which TF is bound by each region. The panels show sensitivity/positive-predictive value curves for the performance obtained using the original (red) and refined motif (green). Sequence logos of the original and refined WMs are shown in each panel as well.

3.9.1 Refining motifs outperforms ordinary motif inference

To check to what extend the refinement of WMs outperforms ordinary motif inference, we ran MEME [39] on the exact same regions that were used by MotEvo for the refinement. For each factor, we either let MEME choose the motif length (green line), that is we set a range ± 5 with respect to the known WM length, or we set it to the length of the known WM (red line). MEME was run such that it was looking for three or fewer motifs. The range for the optimal number of sites was set to $[2, 450]$ (A range that is too small, for example $[2, 50]$, leads to too specialized motifs). Figure 3.10 shows the best sensitivity/positive-predictive value curves, in case several motifs were reported, for the five TFs. In figure 3.11, 3.12, and 3.13 you can see MEME's inferred and MotEvo's refined WMs.

For the NRSF and CTCF data set it simply is enough to use MEME to infer the WMs, even though MEME's WM for CTCF performs a little bit worse on regions containing only weak(er) sites. The reason for that, we believe, is that MEME focuses too much on regions that have a strong site. Therefore, the inferred motif is biased to a subset of regions. When looking at the motifs, it can be seen that MEME's inferred WMs have higher information content (see figure 3.11).

In the case of SRF, MEME infers a reasonable WM as well. Even though MotEvo's and MEME's WMs look quite similar (see figure 3.12), MotEvo's refined WM performs better.

For STAT1 and GABP, the difference in performance is biggest. In the case of GABP, the inferred WMs look similar, but the performance is clearly worse than for MotEvo's refined WM. While MEME and MotEvo are in agreement on the core of the motif (GGAAG, position 6 to 10, GABP - MotEvo), they look different in the other positions, resulting in the different sensitivity/positive-predictive value curves. As for STAT1, MEME was not able to infer the STAT1 WM as known by the literature. By increasing the range for the motif length to $[0, 50]$ a motif was inferred with length 38bsp, which had the characteristic consensus TTCC.GGAA sequence, but this motif could only predict the 57 sequences that it was inferred on.

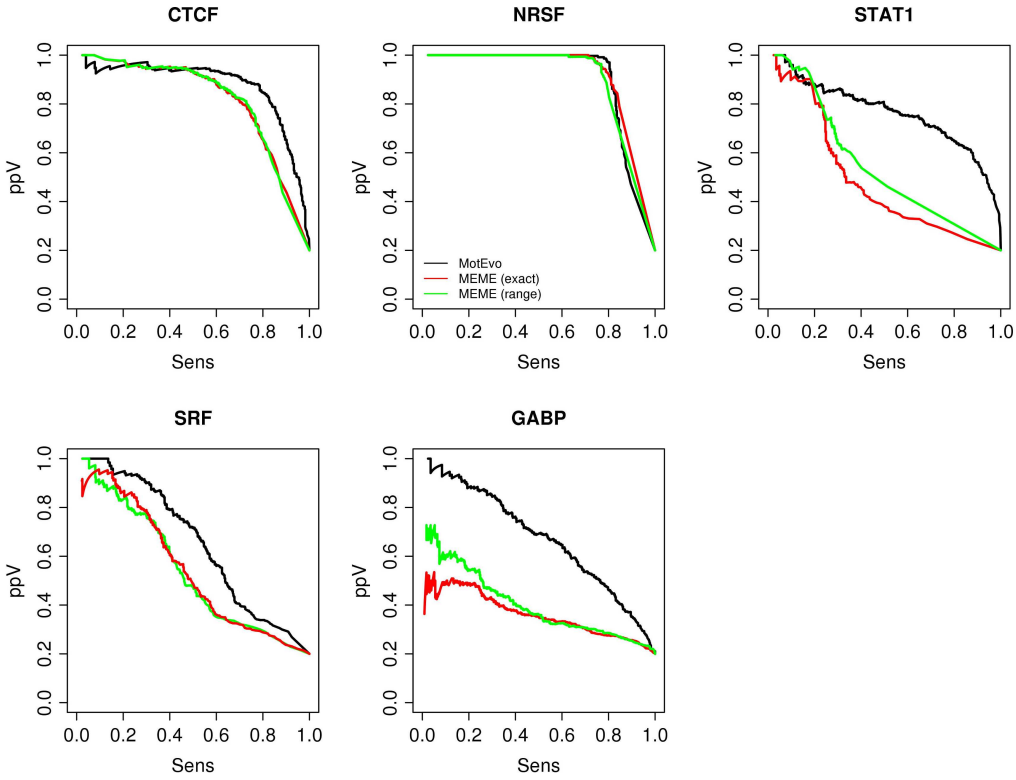


Figure 3.10: For NRSF and CTCF, MEME could successfully infer the WM, which performs equally well as MotEvo's refined WM. For SRF and GABP, there is a clear advantage of the refinement over the inference. Even though the MEME's STAT1 motif seems to do reasonable well at the high signal end, it failed to actually find the motif seen in the literature.

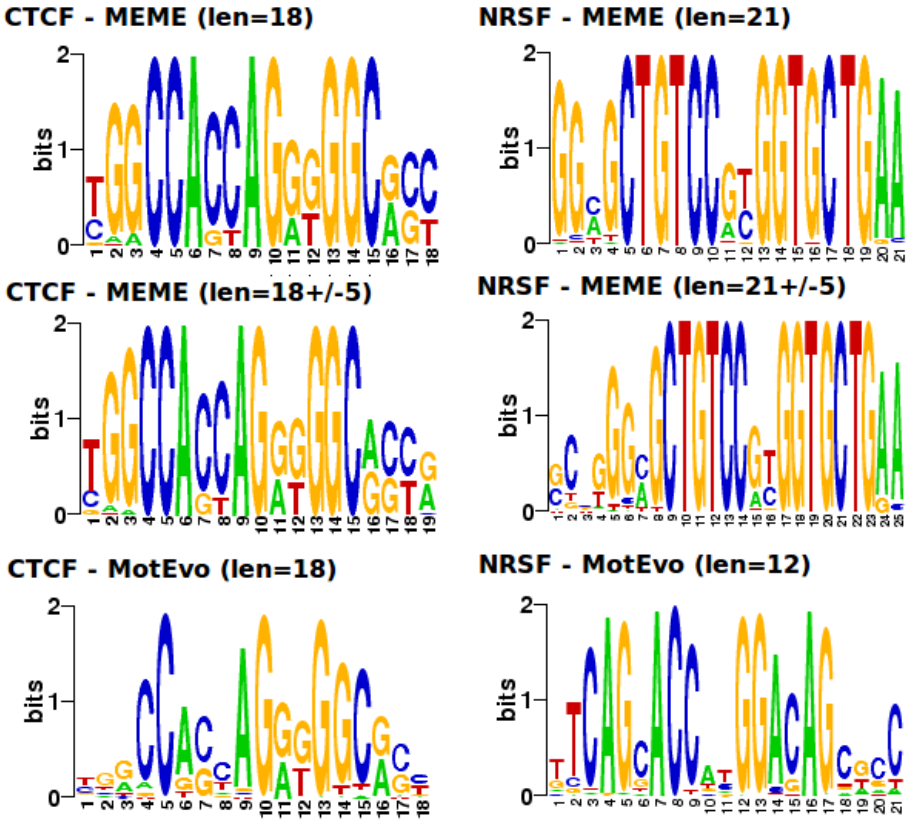


Figure 3.11: MEME's inferred WMs look similar to the one MotEvo refined. The information content of MEME's WM are higher than MotEvo's.

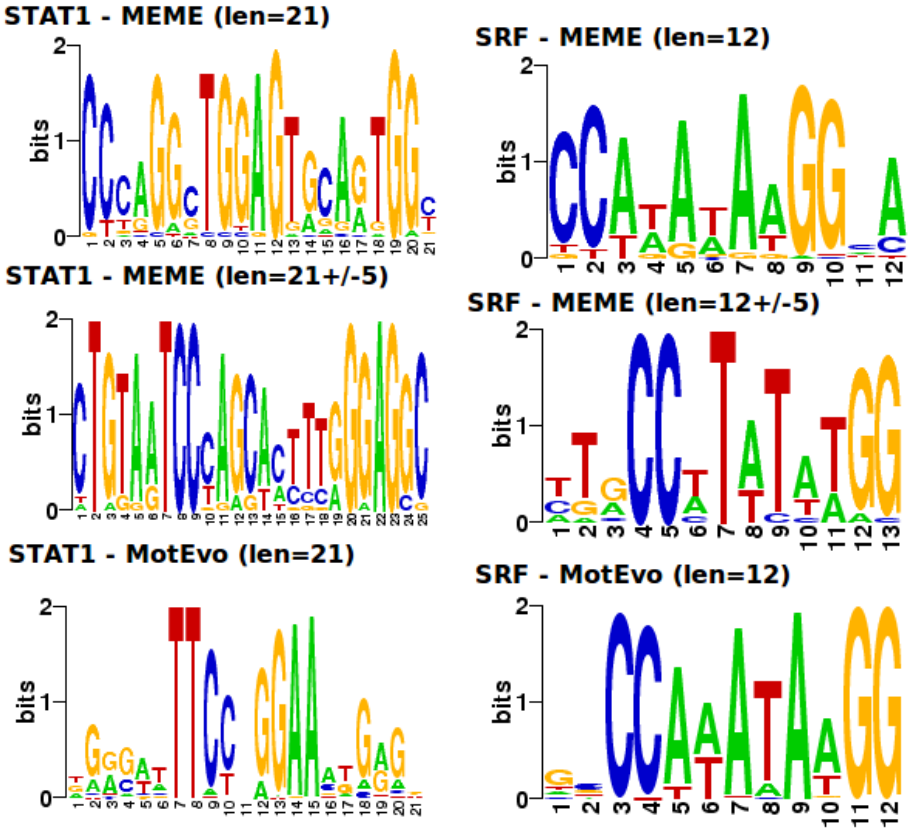


Figure 3.12: MEME failed to infer the STAT1 motif. On the other hand, it inferred the correct motif for SRF, which performs worse than MotEvo's refined version, though.

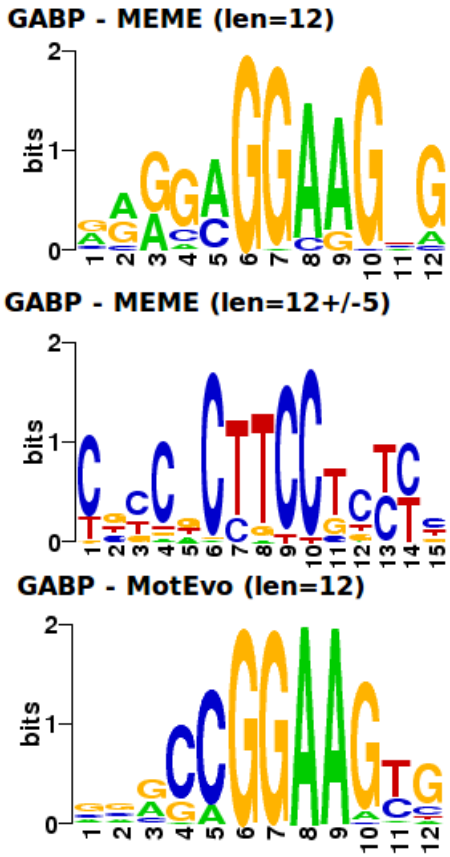


Figure 3.13: MEME inferred the correct motif for GABP, which performs worse than MotEvo's refined version, though.

3.10 Enhancer prediction

To test MotEvo's enhancer predictions we used 76 experimentally validated blastoderm cis-regulatory modules (CRMs) from the REDfly database [40]. For each CRM 5kb of flanking sequence both upstream and downstream were extracted, orthologous sequences from the other Drosophila genomes were obtained, and the sequences were multiply aligned using T-coffee [28]. To remove other (unknown) cis-regulatory modules from these flanking regions we shuffled the columns of the multiple alignments while preserving their gap patterns and their conservation, i.e the subset of species with conserved nucleotides relative to the reference melanogaster sequence.

We then constructed several pruned versions of the multiple alignments by retaining only the sequences from particular subsets of 1, 2, 3, 5, 7, and 9 species. The subsets of species were chosen so as to distribute them roughly uniformly across the tree. For the single species case we of course used only the reference melanogaster sequence. For the two-species case we used melanogaster and pseudo-obscura. For the other cases the specific subsets of species and their subtrees are shown in Fig. 3.14.

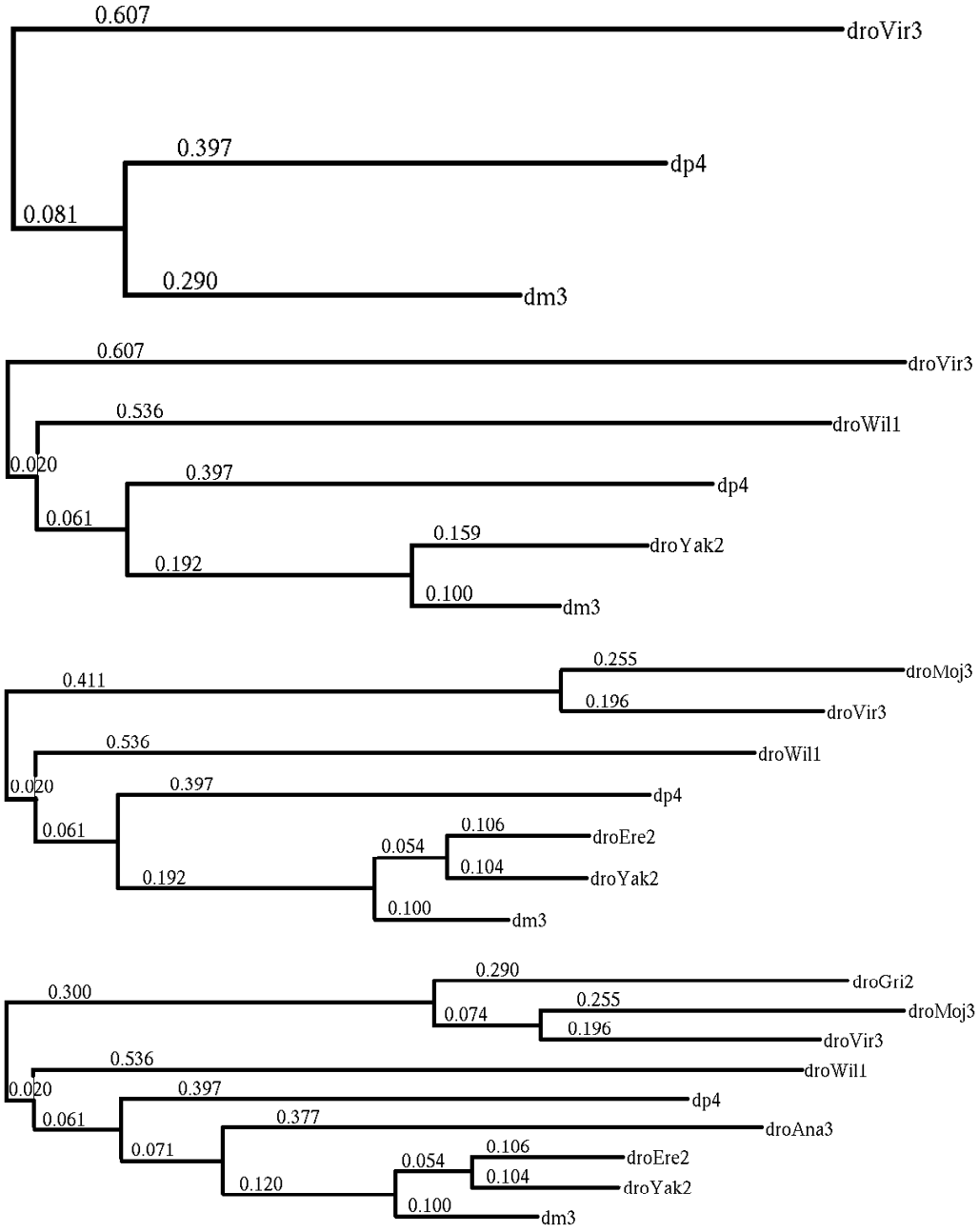


Figure 3.14: Phylogenetic trees of the subsets of Drosophila species with 3, 5, 7, and 9 species, respectively, that were used in the enhancer predictions. Branch lengths are given in expected number of substitutions per neutrally evolving site.

3.11 Comparison to MONKEY and PhyloScan

We compared MotEvo to two TFBS prediction algorithms, namely MONKEY [4] and PhyloScan [32, 33]. For MONKEY (version 2.0), we pruned the original phylogenetic tree in newick format for each region such that it contained only species that occurred in the multi-alignment. The pruned tree was formatted using provided scripts to work with MONKEY. The same nucleotide frequencies for the background model and the same WMs were used for running MONKEY as in the case of MotEvo. For each region, MONKEY reports TFBSs and corresponding p-values. Either the smallest p-value (dotted black line) for having a site or the sum of the log of the p-values for all sites in this region (solid black line) was used to assign a score to each region.

For PhyloScan, we used the online version. Again, the same phylogenetic tree was used as before. Our multi-aligned regions were converted into MAF format. Each WM was represented by a collection of $10,000$ sequences drawn from each WM. For each five WMs, all positions (fragmentation mask) were included. We used four combinations for the site rank weights ($[0.9, 0.1]$, $[0.1, 0.9]$, $[0.5, 0.5]$, $[0.5, 0.25, 0.25]$), all resulting in similar sensitivity/PPV curves. For each region, the reported p-value for the region containing at least a site (solid red line) or the summed log of the e-value for having a site (dotted red line) was used to assign a score.

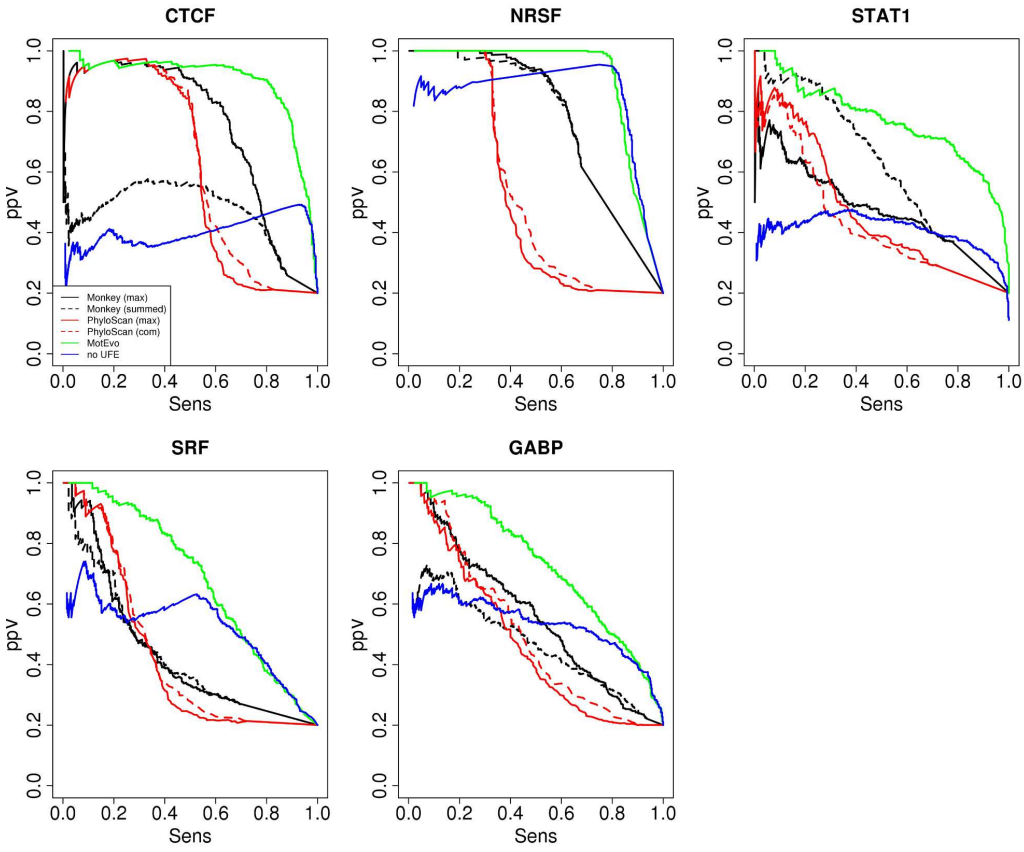


Figure 3.15: Here we show the sensitivity/PPV curves for all five factors. The curves clearly show that MONKEY, PhyloScan, and MotEvo perform equally well for regions with a strong signal, whereas MONKEY and PhyloScan perform worse on regions that seem to contain only weaker sites. In the case of PhyloScan, it seems that under the assumed probabilistic model, it is almost impossible for a weak site to obtain a small p-value. For the sake of completeness, we also show MotEvo's performance when not including the UFE model. As you can see, all the regions that are well conserved but for which the reference species has only a moderate WM score (supplementary Fig. 3.7), result in many falsely predicted regions.

3.12 Dependence of MotEvo's TFBS predictions on different aligners

To test the effect of different alignment methods, we compared MotEvo's TFBSs prediction when using two additional aligners, namely FSA [41] (version 1.15.6) and MUSCLE [31] (version 3.8.31).
Using pairwise genome alignments from the UCSC data-base [27], we extracted orthologous regions from six other mammals (mouse, dog, cow, monkey, horse, and opossum). These orthologous regions were given to the aligners to obtain multiply-aligned regions using default options. Finally, MotEvo was run on the alignments created by T-coffee, FSA, and MUSCLE, and sensitivity/PPV curves were obtained.

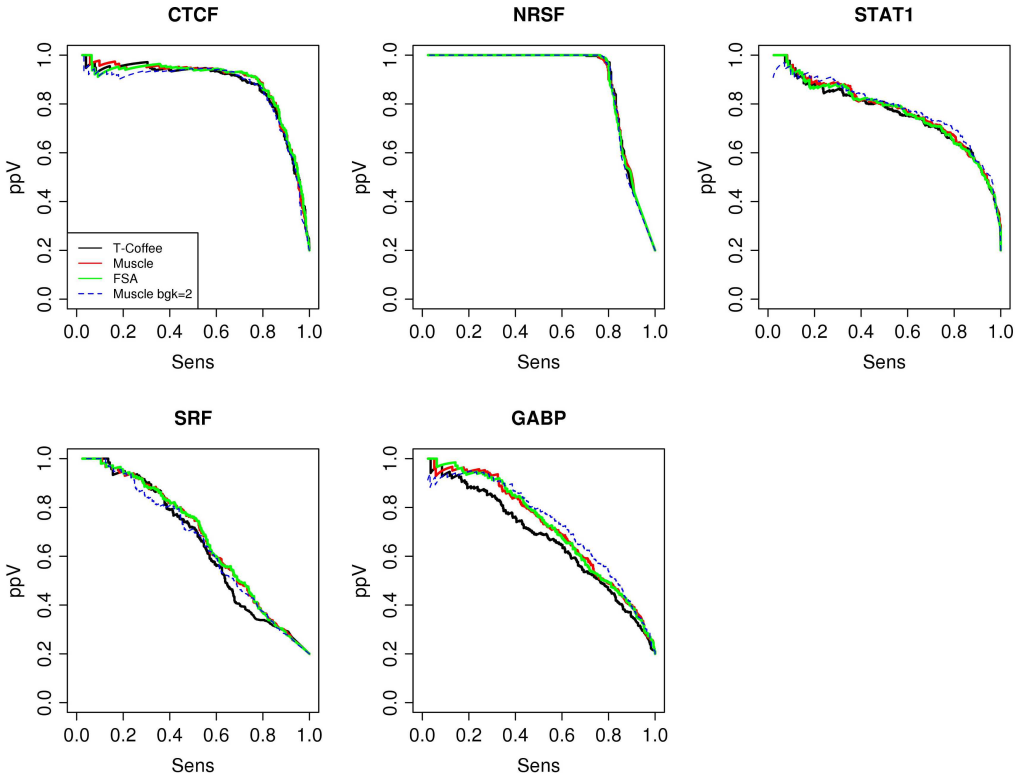


Figure 3.16: All three alignment methods, T-Coffee, FSA, and MUSCLE, result in similar sensitivity/PPV curves. For the sake of completeness, we also ran MotEvo using different Markov orders for the background model (1 and 2). The blue dotted line shows the sensitivity/PPV curve for a 2^{nd} order Markov model. The curve for a 1^{st} order Markov model looks almost identical.

Bibliography

[1] M. L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol.*, 5:201, 2003.

[2] S. Hannenhalli. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, 24:1325–1331, Jun 2008.

[3] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.

[4] A. M Moses, D. Y Chiang, D. A. Pollard, V. N. Iyer, and M. B. Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, 5:R98, 2004.

[5] R. Siddharthan, E. D. Siggia, and E. van Nimwegen. Phylogibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 2005.

[6] J. Hawkins, C. Grant, W. S. Noble, and T. L. Bailey. Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, 25:i339–347, Jun 2009.

[7] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[8] T. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

[9] S. Sinha, M. Blanchette, and M. Tompa. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, 2004.

[10] Eric H. Davidson. *Genomic regulatory systems*. Academic Press, San Diego, 2001.

[11] M. C. Frith, U. Hansen, and Z. Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889, 2001.

[12] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic cis-regulatory modules, applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, 3(30), 2002.

[13] H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23:134–141, Jan 2007.

[14] T. Wasson and A. J. Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, 19:2101–2112, Nov 2009.

[15] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 suppl. 1:i292–i301, 2003.

[16] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, Oct 1990.

[17] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[18] A. L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 5(7):910–917, 1998.

[19] N. Molina and E. van Nimwegen. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Research*, 18:148–160, 2008.

[20] A. Siepel, G. Bejerano, J. S. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15:1034–1050, 2005.

[21] R. Durbin, S. Eddy, G. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[22] E. van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8(Suppl 6):S4, 2007.

[23] David N. Arnosti and Meghana M. Kulkarni. *Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?* Wiley Subscription Services, Inc., A Wiley Company, 2005.

[24] A. Valouev, D.S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Meth*, 5(9):829–834, 2008.

[25] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research*, 36(16):5221–5231, 2008.

[26] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9:R137, 2008.

[27] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, 36:D773–779, Jan 2008.

[28] C. Notredame, D. Higgins, and J. Heringa. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[29] D. Vlieghe, A. Sandelin, P. J. De Bleser, Kris Vleminckx, W. W. Wasserman, Frans van Roy, and B. Lenhard. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucl. Acids Res.*, 34:D95–D97, 2006.

[30] Eimear E Holohan, Camilla Kwong, Boris Adryan, Marek Bartkuhn, Martin Herold, Rainer Renkawitz, Steven Russell, and Robert White. CTCF genomic binding sites in *Drosophila* and the organisation of the bithorax complex. *PLoS Genet*, 3(7):e112, 07 2007.

[31] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797, 2004.

[32] C. S. Carmack, L. A. McCue, L. A. Newberg, and C. E. Lawrence. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol*, 2:1, 2007.

[33] M. J. Palumbo and L. A. Newberg. Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Res.*, 38:W268–274, Jul 2010.

[34] K. M. McBride and N. C. Reich. The ins and outs of STAT1 nuclear transport. *Sci. STKE*, 2003:RE13, Aug 2003.

[35] A. Ivan, M. S. Halfon, and S. Sinha. Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol.*, 9:R22, 2008.

[36] H. Suzuki, A. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. de Hoon, S. Katayama, K. Schroder, P. Carninci, Y. Tomaru, M. Kanamori-Katayama, A. Kubosaki, A. Akalin, Y. Ando, E. Arner, M. Asada, H. Asahara, T. Bailey, V. B. Bajic, D. Bauer, A. G. Beckhouse, N. Bertin, J. Bjorkegren, F. Brombacher, E. Bulger, A. M. Chalk, J. Chiba, N. Cloonan, A. Dawe, J. Dostie, P. G. Engstrom, M. Essack, G. J. Faulkner, J. L. Fink, D. Fredman, K. Fujimori, M. Furuno, T. Gojobori, J. Gough, S. M. Grimmond, M. Gustafsson, M. Hashimoto, T. Hashimoto, M. Hatakeyama, S. Heinzel, W. Hide, O. Hofmann, M. Hornquist, L. Huminiecki, K. Ikeo, N. Imamoto, S. Inoue, Y. Inoue, R. Ishihara, T. Iwayanagi, A. Jacobsen, M. Kaur, H. Kawaji, M. C. Kerr, R. Kimura, S. Kimura, Y. Kimura, H. Kitano, H. Koga, T. Kojima, S. Kondo, T. Konno, A. Krogh, A. Kruger, A. Kumar, B. Lenhard, A. Lennartsson, M. Lindow, M. Lizio, C. Macpherson, N. Maeda, C. A. Maher, M. Maqungo, J. Mar, N. A. Matigian, H. Matsuda, J. S. Mattick, S. Meier, S. Miyamoto, E. Miyamoto-Sato, K. Nakabayashi, Y. Nakachi, M. Nakano, S. Nygaard, T. Okayama, Y. Okazaki, H. Okuda-Yabukami, V. Orlando, J. Otomo, M. Pachkov, N. Petrovsky, C. Plessy, J. Quackenbush, A. Radovanovic, M. Rehli, R. Saito, A. Sandelin, S. Schmeier, C. Schonbach, A. S. Schwartz, C. A. Semple, M. Sera, J. Severin, K. Shirahige, C. Simons, G. St Laurent, M. Suzuki, T. Suzuki, M. J. Sweet, R. J. Taft, S. Takeda, Y. Takenaka, K. Tan, M. S. Taylor, R. D. Teasdale, J. Tegner, S. Teichmann, E. Valen, C. Wahlestedt, K. Waki, A. Waterhouse, C. A. Wells, O. Winther, L. Wu, K. Yamaguchi, H. Yanagawa, J. Yasuda, M. Zavolan, D. A. Hume, T. Arakawa, S. Fukuda, K. Imamura, C. Kai, A. Kaiho, T. Kawashima, C. Kawazu, Y. Kitazume, M. Kojima, H. Miura, K. Murakami, M. Murata, N. Ninomiya, H. Nishiyori, S. Noma, C. Ogawa, T. Sano, C. Simon, M. Tagami, Y. Takahashi, J. Kawai, Y. Hayashizaki, H. Suzuki, A. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. de Hoon, S. Katayama, K. Schroder, P. Carninci, Y. Tomaru, M. Kanamori-Katayama, A. Kubosaki, A. Akalin, Y. Ando, E. Arner, M. Asada, H. Asahara, T. Bailey, V. B. Bajic, D. Bauer, A. G. Beckhouse, N. Bertin, J. Bjorkegren, F. Brombacher, E. Bulger, A. M. Chalk, J. Chiba, N. Cloonan, A. Dawe, J. Dostie, P. G. Engstrom, M. Essack, G. J. Faulkner, J. L. Fink, D. Fredman, K. Fujimori, M. Furuno, T. Gojobori, J. Gough, S. M. Grimmond, M. Gustafsson, M. Hashimoto, T. Hashimoto, M. Hatakeyama, S. Heinzel, W. Hide, O. Hofmann, M. Hornquist, L. Huminiecki, K. Ikeo, N. Imamoto, S. Inoue, Y. Inoue, R. Ishihara, T. Iwayanagi, A. Jacobsen, M. Kaur, H. Kawaji, M. C. Kerr, R. Kimura, S. Kimura, Y. Kimura, H. Kitano, H. Koga, T. Kojima, S. Kondo, T. Konno, A. Krogh, A. Kruger, A. Kumar, B. Lenhard, A. Lennartsson, M. Lindow, M. Lizio, C. MacPherson, N. Maeda, C. A. Maher, M. Maqungo, J. Mar, N. A. Matigian, H. Matsuda, J. S. Mattick, S. Meier, S. Miyamoto, E. Miyamoto-Sato, K. Nakabayashi, Y. Nakachi, M. Nakano, S. Nygaard, T. Okayama, Y. Okazaki, H. Okuda-Yabukami, V. Orlando, J. Otomo, M. Pachkov, N. Petrovsky,

C. Plessy, J. Quackenbush, A. Radovanovic, M. Rehli, R. Saito, A. Sandelin, S. Schmeier, C. Schonbach, A. S. Schwartz, C. A. Semple, M. Sera, J. Severin, K. Shirahige, C. Simons, G. S. Laurent, M. Suzuki, T. Suzuki, M. J. Sweet, R. J. Taft, S. Takeda, Y. Takenaka, K. Tan, M. S. Taylor, R. D. Teasdale, J. Tegner, S. Teichmann, E. Valen, C. Wahlestedt, K. Waki, A. Waterhouse, C. A. Wells, O. Winther, L. Wu, K. Yamaguchi, H. Yanagawa, J. Yasuda, M. Zavolan, D. A. Hume, T. Arakawa, S. Fukuda, K. Imamura, C. Kai, A. Kaiho, T. Kawashima, C. Kawazu, Y. Kitazume, M. Kojima, H. Miura, K. Murakami, M. Murata, N. Ninomiya, H. Nishiyori, S. Noma, C. Ogawa, T. Sano, C. Simon, M. Tagami, Y. Takahashi, and J. Kawai. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, 41:553–562, May 2009.

[37] K. Chen, E. van Nimwegen, N. Rajewsky, and M.L. Siegal. Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biology and Evolution*, 2:697–707, 2010. PMID:20829281.

[38] B. Wilczynski and E. E. Furlong. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, 6:383, Jun 2010.

[39] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

[40] SM Gallo, DT Gerrard, D Miner, M Simich, B Des Soye, CM Bergman, and MS Halfon. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.*, 39:D118–23, 2011.

[41] R. K. Bradley, A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. Fast statistical alignment. *PLoS Comput. Biol.*, 5:e1000392, May 2009.

Chapter 4

ISMARA: Modeling genomic signals as a democracy of regulatory motifs

Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik van Nimwegen

Manuscript in preparation (January 30, 2013).

Accurate reconstruction of the regulatory networks that control gene expression is one of the key current challenges in molecular biology. Although gene expression and chromatin state dynamics are ultimately encoded by constellations of binding sites recognized by regulators such as transcriptions factors (TFs) and microRNAs (miRNAs), our understanding of this regulatory code and its context-dependent read-out remains very limited. Given that there are thousands of potential regulators in mammals, it is not practical to use direct experimentation to identify which of these play a key role for a particular system of interest.

We developed a methodology that uses genome-wide predictions of TF binding sites and miRNA target sites to model gene expression or chromatin modifications in terms of these sites, and completely automated it into a web-based tool called ISMARA (Integrated System for Motif Activity Response Analysis), located at `http://ismara.unibas.ch`. Given as input only gene expression or chromatin state data across a set of samples, ISMARA identifies the key TFs and miRNAs driving expression/chromatin changes and makes detailed predictions regarding their regulatory roles. These include predicted activities of the regulators across the samples, their genome-wide targets, enriched gene categories among the targets, and direct interactions between the regulators.

Applying ISMARA to data sets from well-studied systems, we show that it consistently identifies known key regulators ab initio. We also present a number of novel predictions including regulatory interactions in innate immunity, a master regulator of mucociliary differentiation, TFs consistently upregulated in cancer, and TFs that mediate specific chromatin modifications.

4.1 Introduction

Since the seminal work of Jacob and Monod [1], much has been learned about the molecular mechanisms by which gene expression is regulated, and the molecular components involved. Historically, most work has focused on transcription factors (TFs), arguably the most important regulators of gene expression, which

bind to cognate sites in the DNA, frequently in the neighborhood of transcription start sites (TSSs), and regulate the rate of transcription initiation. However, more recently it has become clear that the state of the chromatin, which can be modulated through modifications of the DNA nucleobases and of the histone tails of nucleosomes, also plays a crucial role. For example, the local chromatin state affects the ability of TFs to access their binding sites, and the chromatin state can in turn be modified through TF-guided recruitment of chromatin modifying enzymes. Furthermore, an entirely new layer of post-transcriptional regulation has been uncovered in recent years in the form of microRNAs (miRNAs) [2]. These guide RNA-induced silencing complexes to target mRNAs, inhibiting their translation and accelerating their decay [3].

In spite of these many insights, our current understanding of the function of genome-wide gene regulatory networks in mammals is still rudimentary. For example, we only know the sequence specificity of less than 500 [4–6] of the approximately 1500 [7] TFs in mammalian genomes. Our knowledge of how TF binding is affected by chromatin state, of the combinatorial interactions between TFs and their co-factors, and the impact of post-translational modifications on TF activity, is even more fragmentary. Our understanding of the transcriptome-wide effects of miRNAs on their targets is similarly limited. It is thus clear that we are still far from being able to develop realistic quantitative models of gene regulatory networks in mammals. Consequently, rather than aiming to develop comprehensive computational models of gene regulatory dynamics, the most constructive contribution that computational approaches can currently provide is to develop models that help guide experimental efforts.

Given a particular mammalian subsystem or process, e.g. a particular developmental or cellular differentiation process, or the response of a tissue to a particular perturbation, the initial steps in unraveling its gene regulatory circuitry are to identify the key regulators in the process, and to characterize the rough functional roles these regulators play. However, for the vast majority of mammalian systems these initial steps have yet to be taken. Given the large number of potential regulators, a direct experimental approach, e.g. through large-scale screening, is typically not feasible. There is thus a strong need for computational methods that, given a system of interest, can predict key regulatory players and make concrete, directly testable, hypotheses about their regulatory roles. We here present an integrated and completely automated computational methodology that accomplishes exactly this task.

Our approach, ISMARA (Integrated System for Motif Activity Response Analysis), capitalizes on a number of recent computational and experimental technological developments. First, whereas large-scale screening of the functions of individual regulators in a particular system is often impractical, it is relatively straight forward to measure gene expression (i.e. with microarray or RNA-seq) or chromatin state (with ChIP-seq) in high-throughput across a set of samples of interest. Second, over the last years sophisticated comparative genomic methods have been developed that allow relatively accurate computational prediction of regulatory sites for hundreds of TFs and miRNAs on a genome-wide scale [8–10]. Third, through extensive experimental efforts, genome-wide annotations of transcript structures [11] and promoters [12] have also become available.

Given as input a set of genome-wide gene expression or chromatin state measurements across a number of samples, ISMARA models the gene expression or chromatin state dynamics in terms of a comprehensive set of computationally predicted regulatory sites, using a simple linear modeling approach called Motif Activity Response Analysis (MARA) that we originally proposed in [13]. As a result, ISMARA identifies the key regulators (i.e. TFs and miRNAs) driving gene expression/chromatin state changes across the samples, the activity profiles of these regulators, their target genes, and the sites on the genome through which these regulators act. The analysis is carried out within a completely automated system, which combines pre-calculated annotations of regulatory sites for hundreds of regulators across promoters in mammalian genomes with processing of input data, automated tuning of parameters, and post-processing to provide a large collection of auxiliary analysis results. To use ISMARA, all that users need to do is upload their data to the web-server `http://ismara.unibas.ch/` and submit it to the system, after which all results are presented through a user-friendly graphical web-interface. Importantly, in ISMARA the motif activity response analysis has been extended to model not only gene expression data from various platforms (microarray, RNA-seq), but

essentially any sequencing data reflecting a genomic mark (ChIP-seq) including chromatin modifications or TF binding. In addition, ISMARA models not only the effect of TFs on mammalian gene expression, but also the effect of miRNAs. Below we will first describe the methodology used by ISMARA and the results that it provides, and then we will demonstrate its power through a number of applications.

4.2 Results

4.2.1 An Integrated System for Motif Activity Response Analysis

The integrated system for motif activity response analysis (ISMARA) that we developed is schematically depicted in Fig. 4.1. Detailed descriptions of all procedures are provided in the supplementary methods. The system capitalizes on two key resources developed in our group (Fig. 4.1A-C). The first is the genome-wide annotation of promoters in human and mouse, i.e. so-called "promoteromes", that we constructed [12] from genome-wide transcription start site data (deepCAGE data [14]). We supplement these promoter sets with 5' ends of known RNA transcripts from human and mouse, and associate transcripts with promoters. The second key resource that we employ is a genome-wide annotation of functional transcription factor binding sites (TFBSs) that we obtained with Bayesian probabilistic methods for quantifying evolutionary selection pressure, which we developed previously [8, 10]. Briefly, we constructed multiple alignments of orthologous proximal promoter regions across 7 mammalian genomes and curated a collection of approximately 200 non-redundant mammalian regulatory motifs (positional weight matrices) that represent the DNA binding specificities of close to 350 TFs in both human and mouse. We then used our MotEvo algorithm [10] to predict functional TFBSs for all TF regulatory motifs across all promoters in human and mouse (Fig. 4.1A,C). MotEvo is a Bayesian algorithm which explicitly models the evolution of TFBSs across the mammalian phylogeny (Suppl. Fig. 1).

When modeling expression data, ISMARA also integrates the effects of miRNAs that increase decay of transcripts by binding to sites that are generally located in the 3' untranslated regions (UTRs) of transcripts. We used miRNA target site predictions from TargetScan using preferential conservation scoring (P_{CT}) [9], and calculated an overall score for the targeting of a promoter by a particular miRNA by averaging over all transcripts associated with the promoter (Suppl. Methods).

The result of the regulatory site annotation was, for both human and mouse, a large matrix $\mathbf{N}$, where N_{pm} is the predicted total number of functional binding sites in promoter p for motif m, where m runs over the 190 TF binding motifs as well as the 86 miRNA 'seed' motifs.

The next step in ISMARA consists of the construction of a data matrix $\mathbf{E}$, where E_{ps} denotes the 'signal' associated with promoter p for sample s. When provided gene expression data in the form of microarrays, ISMARA applies standard normalization procedures and maps the probes on the microarray to the set of known RNA transcripts, which are each in turn associated with promoters. Microarray platforms currently supported by ISMARA are listed in Suppl. Table 1. For a promoter p, the expression E_{ps} is given by the average log-intensity in sample s of the probes associated with promoter p. Similarly, for RNA-seq data the reads are mapped to the known RNA transcripts and E_{ps} is calculated as the average of the logarithm of the fraction of all reads in the sample that map to transcripts associated with promoter p. When processing ChIP-seq data, the signal E_{ps} is calculated as the logarithm of the fraction of reads in sample s that map to a 2 kilobase region centered on promoter p. Details of the normalization steps involved are again provided in the supplementary methods.

At the core of ISMARA is the MARA model [13] which, similar to previous linear modeling approaches [15, 16], assumes that the 'signal' at each promoter p is a linear function of its binding sites N_{pm}:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \tag{4.1}$$

where c_p is a term reflecting the basal activity of promoter p, $\tilde{c}_s$ reflects the total expression in sample s,

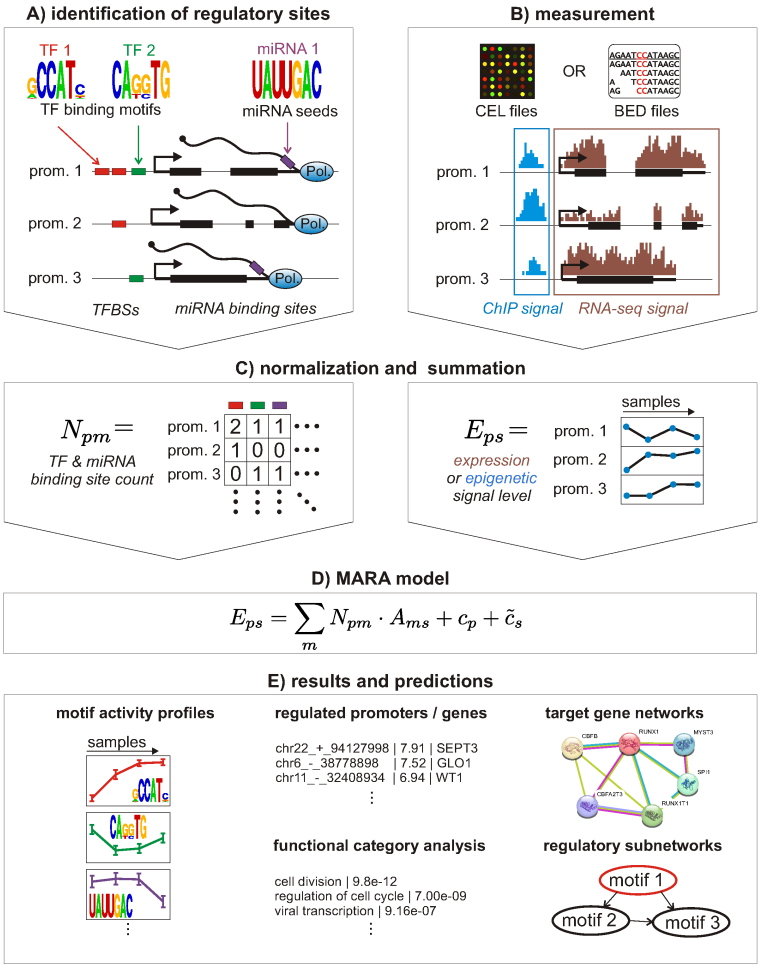


Figure 4.1: Outline of the Integrated System for Motif Activity Response Analysis (ISMARA). **A:** ISMARA starts from a curated genome-wide collection of promoters and their associated transcripts. Using a comparative genomic Bayesian methodology [10], transcription factor binding sites (TFBSs) for $\approx$ 200 regulatory motifs are predicted in proximal promoters. Similarly, miRNA target sites for $\approx$ 100 seed families are annotated in the 3' UTRs of transcripts associated with each promoter. **B:** Users provide measurements of gene expression (microarray, RNA-seq) or chromatin state (ChIP-seq). The raw data are processed automatically and, for each promoter and each sample, a signal is calculated. For ChIP-seq data, the signal is calculated from the read density in a region around the transcription start. For gene expression data, the expression signal is calculated from read densities across the associated transcripts (RNA-seq) or intensities of associated probes (microarray). **C:** The site predictions and measured signals are summarized in two large matrices. The components N_{pm} of matrix $\mathbf{N}$ contain the total number of sites for motif m (TF or miRNA) associated with promoter p. The components E_{ps} of matrix $\mathbf{E}$ contain the signal associated with promoter p in sample s. **D:** The linear MARA model is used to explain the signal levels E_{ps} in terms of bindings sites N_{pm} and unknown motif activities A_{ms}, which are inferred by the model. The constants c_p and $\tilde{c}_s$ correspond to basal levels for each promoter and sample, respectively. **E:** As output, ISMARA provides the inferred motif activity profiles A_{ms} of all motifs across the samples s, sorted by the significance of the motifs. A sorted list of all predicted target promoters is provided for each motif, together with the network of known interactions between these targets (provided by the String database, http://string-db.org/), and a list of Gene Ontology categories that are enriched among the predicted targets. Finally, for each motif, a local network of predicted direct regulatory interactions with other motifs is provided.

and A_{ms} is the (unknown) *activity* of motif m in sample s. That is, using the predicted site counts N_{pm} and the experimentally measured E_{ps}, we use the model (4.1) to infer the activities A_{ms} of all motifs across all samples. To infer the activities, ISMARA uses a Bayesian procedure with a Gaussian likelihood model for the difference between the measured signal E_{ps} and the predicted signal, and a Gaussian prior distribution for the activities (Methods). The latter is used to avoid overfitting and ISMARA uses a cross-validation procedure to set the parameters of the prior (see Suppl. methods). The entire posterior distribution of motif activities is a multi-variate Gaussian which is determined using singular value decomposition (see Suppl. methods).

It is important to note that we do not expect the simple model (4.1) to provide an accurate fit to the signal E_{ps} at individual promoters. As mentioned in the introduction, many factors that influence expression and local chromatin state are not included in our model. Moreover, instead of each binding site contributing linearly to E_{ps}, in reality the expression E_{ps} will likely be a complex combinatorial function of the constellation of binding sites in promoter p. Indeed, we typically find that the simple model (4.1) captures only a small fraction of the variance of E_{ps} across the samples (Suppl. Fig. 2). However, the aim of the model (4.1) is not to fit the signals E_{ps}, but rather to identify which of the motifs m play an important role, and how these motifs contribute to E_{ps} across the samples. Since each motif m targets hundreds to thousands of promoters p, the inferred motif activities A_{ms} are statistical averages of the behaviors of a large number of promoters. This averaging causes the complexities at individual promoters to effectively cancel out and ensures that the overall influence of a motif can still be reliably inferred. To put it differently, if a clear average contribution of a given motif m is detected using the simple linear model (4.1) in spite of it being a poor model at individual promoters, we can be confident that the motif indeed contributes to the signal E_{ps}.

Apart from inferring motif activities, ISMARA also predicts which individual promoters are regulated by each motif m. As detailed in the Suppl. methods, for each promoter with predicted TFBSs for the motif (i.e. $N_{pm} > 0$) ISMARA estimates the log-likelihood ratio S_{pm} of the entire model with the TFBSs for m in p present, and the model in which the entry N_{pm} has been set to zero. That is, S_{pm} rigorously quantifies how much removal of the sites for m in p decreases the fit of the model to the data.

4.2.2 Overview of the results presented by ISMARA

We have made ISMARA available through a web interface http://www.ismara.unibas.ch as part of our Swiss-Regulon resources [6]. Users can directly upload unprocessed microarray (CEL files), RNA-seq, or ChIP-seq data (bed files) which are then analyzed automatically without the need for any additional input from the user (Fig. 4.1B). The results are made available through a web interface and can also be downloaded in flat-file format. To give an overview of the results ISMARA provides, we applied it to the GNF Gene Atlas [17] of mRNA expression profiles across 91 tissues and cell lines in mouse. The results are available at http://ismara.unibas.ch/supp/dataset1/ismara_report/.

The first output of ISMARA is a list of all regulatory motifs sorted by a z-score which summarizes the importance of the motif for explaining the expression variation across the samples. This score roughly corresponds to the average number of standard-deviations the motif activity is away from zero (see Methods and Suppl. Methods). Besides the z-score of each motif, the list also displays the set of associated TFs, a thumbnail of its activity across the input samples, and a sequence logo for each motif (Suppl. Fig. 3). In the Gene Atlas data, the second most significant motif is E2F1..5, corresponding to the E2F1 through E2F5 transcription factors that are known to regulate the cell cycle [18,19]. Following the link from the motif name links leads to a page with additional details regarding the E2F1..5 motif (Suppl. Figs. 4-6), including its inferred activity profile across the samples, once ordered according to the user's input (Suppl. Fig. 4), and once ordered according to the sample-dependent activity z-values (Suppl. Fig. 5). The samples in which the E2F activity is highest are known to be composed of fast dividing cells (bone marrow, hematopoietic stem cells and artificial cell lines), while neural tissues, containing largely non-dividing cells have the lowest E2F activity (Fig. 4.2A). The page also provides a list of predicted target promoters of the motif, sorted by their score S_{pm} (Suppl. Fig. 6). Besides the score, this list includes for each target a link to a genome browser

view of the promoter which shows the predicted TFBSs (Suppl. Fig. 7), associated gene and transcripts, and a short description of each target gene. The user can interactively change how many of the top targets are shown, or search for a gene or transcript of interest in the list of all targets. To provide the user with a more intuitive picture of the predicted list of targets of the motif, a link is provided to a network view of the target genes as provided by the STRING database [20], where network links indicate known functional associations between the genes. For E2F1..5, the STRING network reveals a large, highly connected cluster of predicted targets that are known to be involved in cell cycle, and particularly in DNA replication (Suppl. Fig. 8). The role of the E2F1..5 motif in the cell cycle is further confirmed by Gene Ontology analysis [21] which shows that DNA replication, S phase, and regulation of DNA replication are categories whose genes are most highly enriched among the targets of E2F (Suppl. Fig. 9). Thus, based only on expression data, MARA predicts E2F to be a key regulator of cell proliferation, with E2F activity acting effectively as a marker for proliferation.

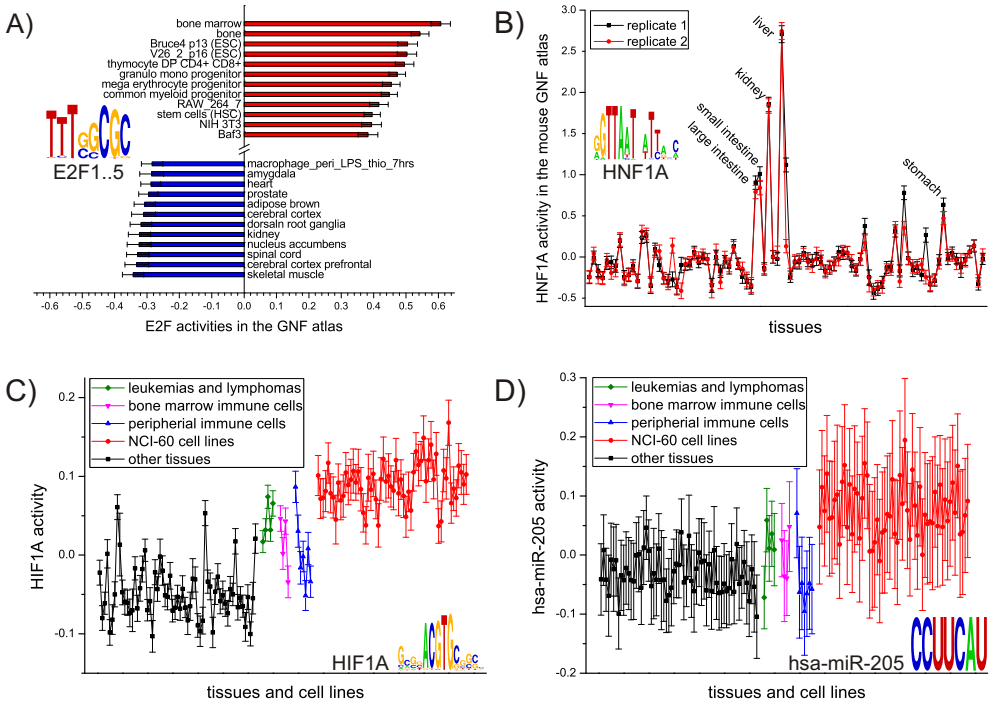


Figure 4.2: Motif activities in the GNF Gene Atlas. **A**: Tissues showing the highest and lowest activities of the E2F motif (shown as inset). **B**: Activity of the HNF4A/NR2F1,2 motif (shown as inset) across all tissues for the two biological replicates (red and black lines). Names of the tissues with highest activity are indicated. **C**: Activity of the HIF1A motif (shown as inset) across the human GNF and NCI-60 samples. Subsets of samples corresponding to NCI-60 cancer cell lines (red), leukemia and lymphomas (green), peripheral immune cells (blue), bone marrow immune cells (pink), and all other tissues (black) are shown in different colors. **D**: As for panel C, but now for the miRNA hsa-miR-205 (seed sequence shown as inset).

For many of the regulatory motifs there are multiple TFs that can bind to the sites of the motif and it is not *a priori* clear which of the TFs is most responsible for the motif activity in a given system. Note that the motif activity is inferred from the behavior of the predicted *targets* of the motif. That is, roughly speaking, an increasing activity is inferred when its targets show on average an increase in expression, that cannot be explained by the presence of other motifs in their promoters. The mRNA expression profiles of the TFs associated with a motif thus provide independent information about the link between the TFs and

the motif activities, and ISMARA provides an analysis of the correlation between motif activities and the expression profiles of the associated TFs for each motif. For example, for the case of E2F1..5, the expression of all associated TFs except E2F5 show a very significant positive correlation with the motif activities (Suppl. Figs. 10 and 11). This also shows that these TFs act as *activators*. That is, whenever a negative correlation between motif activity and TF expression is observed, the TF most likely acts as a repressor, e.g. as observed for the known repressor REST (Suppl. Fig. 12). However, it should be noted that motif activity does not need to be a direct function of TF expression, i.e. the effect of a TF on its targets will not only depend on its expression, but possibly on post-translational modifications, on cellular localization, and on the presence of specific co-factors. Therefore, although a strong correlation between TF expression and motif activity is a good indication that the TF is responsible for the motif activity, the absence of such a correlation does not imply that the TF is not involved in the motif's activity.

To gain insight in the transcription regulatory networks that control expression profiles, it is of particular interest to identify direct regulatory connections between the TFs themselves. In ISMARA, a predicted transcription regulatory interaction from motif m to m' occurs when motif m is predicted to target a promoter of one of the TFs associated with m'. To visualize predicted motif-motif regulatory networks ISMARA provides, for each motif m, a local network picture that shows all predicted regulatory connections between m and other regulatory motifs. The user can interactively change the cut-off on the target score S_{pm} to draw this picture. For E2F1..5 we find that the strongest predicted targets are the promoters of Myb, of TFDP1, and of the E2F2 gene (Suppl. Fig. 13). Indeed, the c-Myb promoter is known to be regulated by an E2F site [22] and the E2F2 promoter has indeed been shown to be bound directly by E2F4 [23]. The transcription factor TFDP1 forms hetero-dimers with various members of the E2F family and ISMARA predicts that this co-factor of the E2F family is itself regulated through an E2F site. To our knowledge, this is novel prediction.

An example of a motif with highly condition-specific activity is HNF1A (Fig. 4.2B). The associated transcription factor hepatocyte nuclear factor 1 homeobox A is relatively well-studied and known to be mainly expressed in liver, kidney, stomach and intestine [24,25], where it is essential for organ function [26]. Indeed, ISMARA infers that the HNF1A activity is by far the highest in liver and kidney, followed by intestinal tissues and stomach. In addition to its role in these tissues, HNF1A has also been shown to be important for the function of pancreatic islets, and HNF1A mutations causes monogenic diabetes [25]. Indeed, ISMARA predicts high activity for HNF1A in pancreas as well, where its activity ranks 6th and 7th among all motifs in the two replicate samples (Suppl. Fig. 14). Figure 4.2B also illustrates that the inferred motif activities are highly reproducible, in fact more reproducible than the expression profiles from which the motif activities were inferred (Suppl. Fig. 15). The reason for this high reproducibility is that motif activity is inferred from the statistics of all (typically hundreds) of its target promoters.

Experiments are often performed in multiple replicates and one would typically be specifically interested in those motifs that behave reproducibly across the replicates. To this end the ISMARA results page links to a section where users can provide replicate annotation for their samples, which than enables ISMARA to calculate motif activity profiles that are averaged over replicates using a rigorous Bayesian procedure (see Suppl. Methods). As an example, the replicate-averaged results for the mouse GNF atlas are available at http://ismara.unibas.ch/supp/dataset1/averaged_report/.

Apart from averaging over replicates, this procedure can also be used to calculate contrasts between subsets of samples. To illustrate this, we jointly analyzed the human GNF atlas of 79 tissues and cell lines [27] and the NCI-60 reference cancer cell lines [28] (full results at http://ismara.unibas.ch/supp/dataset2/ismara_report/). By treating all non-tumor samples as one condition and all tumor samples as another condition in the averaging, we can identify motifs that are consistently dis-regulated in cancer. Supplementary tables 2 and 3 show the motifs that are most consistently up-regulated or down-regulated in tumors. Among the top up-regulated motifs are several key transcriptional regulators that are well known in cancer biology such as Hif1a [29] (Fig. 4.2C), Myc [30], and E2F [31]. ISMARA also identifies a number of miRNAs whose targets are either consistently upregulated, e.g. miR-205 (Fig. 4.2D) and miR-26, or consistently down-regulated, e.g. miR-24 and the miR-17/93/106 seed family, in tumors. Indeed, multiple studies have

found miR-205 to be down-regulated in a number of different cancers, and miR-205 has been shown to have tumor suppressor function [32–36]. It has also been shown that miR-26a delivery suppresses hepatic tumors in mouse [37], supporting the downregulation of this miRNA in cancer. Similarly, miR-17 is a known oncogene [38], supporting that its targets are down in cancer. The literature on miR-24 function in cancer is more ambiguous [39]. Some evidence has been provided that miR-24 acts as repressor of apoptosis and is upregulated in certain cancers [40]. On the other hand, another study found that miR-24 can inhibit proliferation [41]. Notably, the latter study suggested that miR-24 acts through seedless target sites, which by construction are not detected by TargetScan. In summary, in this system ISMARA successfully identified oncogenes and tumor suppressors *ab initio*.

4.2.3 Inferring motif activity dynamics: inflammatory response

To illustrate ISMARA's analysis of time series data, we applied it to a time series of expression data obtained after activation of human umbilical vein endothelial cells (HUVECs) with tumor necrosis factor (TNF, previously also known as TNFα). Messenger RNA expression was measured every 15 minutes for the first 4 hours after treatment, and every 30 minutes for the next 4 hours [42]. Whereas the original study focused solely on nascent transcription, standard application of ISMARA to this data set (http://ismara.unibas.ch/supp/dataset3/ismara_report/) uncovers the transcription regulatory network involved in this inflammatory response in remarkable detail.

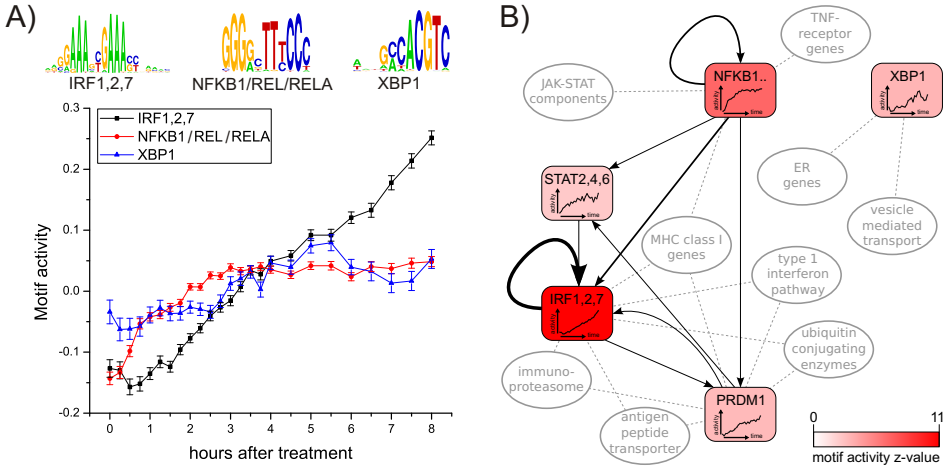


Figure 4.3: Analysis of an inflammatory response time series of human umbilical vein endothelial cells responding to TNF. **A**: Time-dependent activities of the 3 most significant motifs, i.e. NFκB (red), IRF1/2 (black), and XBP1 (blue). Error-bars denote uncertainties in the inferred activities. **B**: Summary of the inferred core regulatory network. Selected top motifs are shown together with interactions between them and pathways/functional categories that are enriched among the targets of these motifs. The intensity of the color corresponds to the z-score of the motif, its time-dependent activity is indicated inside the node, and the thickness of each edge corresponds to its target score S_{pm}.

The response of endothelial cells to TNF is known to be mediated by the NFκB, GATA2, IRF1, and AP-1 [43] TFs. NFκB in particular is crucial for the resulting inflammatory response [44]. Indeed, ISMARA infers that the two most significant motifs are IRF1,2,7 and NFκB. The activity of NFκB increases sharply in the first 45 minutes and slower afterwards, until it reaches a steady activity after 3 hours. The activity of the IRF1,2,7 motif increases steadily starting at 30 to 45 minutes after treatment until the end of the time course (Fig. 4.3A). As shown by NFκB's local network figure (Fig. 4.3B and on the ISMARA results website),

ISMARA infers that IRF1 is activated directly at the level of transcription by NFκB, which is indeed known from previous studies [45]. Other predicted targets of NFκB that are also found to be significantly upregulated in this process are TNF receptor genes, components of the JAK-STAT pathway (note that STAT2,4,6 is the 11th most significant motif, indicating that STAT activity changes, affecting the level of *its* targets) and MHC class I genes. The latter are also predicted to be regulated by IRF1,2,7, which is confirmed by experimental data [46]. ISMARA also predicts that both NFκB and IRF1,2,7 activate the 5th most significant motif, PRDM1 (BLIMP-1), which is an important developmental regulator in the B-cell and T-cell lineages and is required for the secretory pathway in B-cells [47]. PRDM1 activity increases, like that of IRF, across the entire time course, and these two TFs appear to share many of their predicted targets, including type 1 interferon pathway genes, the immuno-proteasome [48], ubiquitin conjugating enzymes, antigen peptide transporters, and MHC class I genes. All of these targets are consistent with the activation of the antigen presenting pathway by these TFs. Finally, the 3rd most significant motif is XBP1, which is activated only after 2.5 hours. Its predicted targets are highly over-represented for endoplasmic reticulum (ER) genes and genes involved in vesicle-mediated and Golgi transport, consistent with the fact that XBP1 is a major regulator of ER stress and the unfolded protein response (UPR) [49]. Moreover, several studies support that the UPR is a general characteristic resulting from inflammation or TNF activation in endothelial cells [50, 51]. Interestingly, the induction of XBP1's activity occurs at the same time as the NFκB activity stops increasing which is in line with studies showing that the UPR can attenuate NFκB induction of inflammation [52–54]. All these predictions of ISMARA, which were made *ab initio* using only the time course expression data, are summarized in the network picture Fig. 4.3B.

Finally, the induction of XBP1's activity is not reflected in the expression of XPB1 itself, which is almost constant across the time course (Suppl. Fig. 16). This underscores that ISMARA infers a motif's activity from the expression of its predicted targets and does not use the regulator's own expression. Indeed, it has been established that XBP1 activity is regulated post-transcriptionally through alternative splicing [55, 56].

4.2.4 Identifying novel master regulators: Mucociliary differentiation of bronchial epithelial cells

Next, we turned to an example system for which much less is known, namely the mucociliary differentiation of bronchial epithelial cells on an air-liquid interface. Aiming to elucidate the regulation of bronchial development, Ross et al. [57] performed differentiation experiments in triplicate over a period of 28 days with cells from three separate donors. This data was then analyzed with commonly used bioinformatic procedures, i.e. genes were clustered into co-expression clusters, and the clusters were analyzed for over-represented gene ontology categories and pathways. This analysis uncovered clusters associated with TGFβ pathway genes, extra-cellular adhesion genes, and genes associated with the microtubule cytoskeleton, but no key regulators or regulatory interactions that drive these expression changes were identified.

In contrast, applying ISMARA to this gene expression data set, we obtain the prediction that by far the most important regulatory motif in this system is RFX, whose activity is strongly increasing over the period from roughly day 4 to day 10 in all 3 donors (Fig. 4.4A, http://ismara.unibas.ch/supp/dataset4/ismara_report/). The predicted targets of RFX are highly enriched in genes known to be associated with cilium assembly, axoneme, and the microtubule cytoskeleton genes (Fig. 4.4B) suggesting that RFX directs ciliogenesis in bronchial epithelial cells.

The RFX family of TFs contains 7 members and it is not *a priori* clear which of these are driving the bronchial differentiation. Comparison of the mRNA expression profiles with activity profiles shows that two of the family members, RFX2 and RFX3 exhibit a striking correlation in their expression with the motif activity (Fig. 4.4A and C). Together these results strongly suggest that the TFs RFX2/3 are master regulators of ciliogenesis in this system. This prediction is consistent with previous studies that have shown that RFX3 is necessary for the ciliogenesis of nodal cilia in mouse embryonic development [58] and during ciliogenesis of motile cilia in a mouse cell-culture system [59]. More specifically, in the latter study it was found that

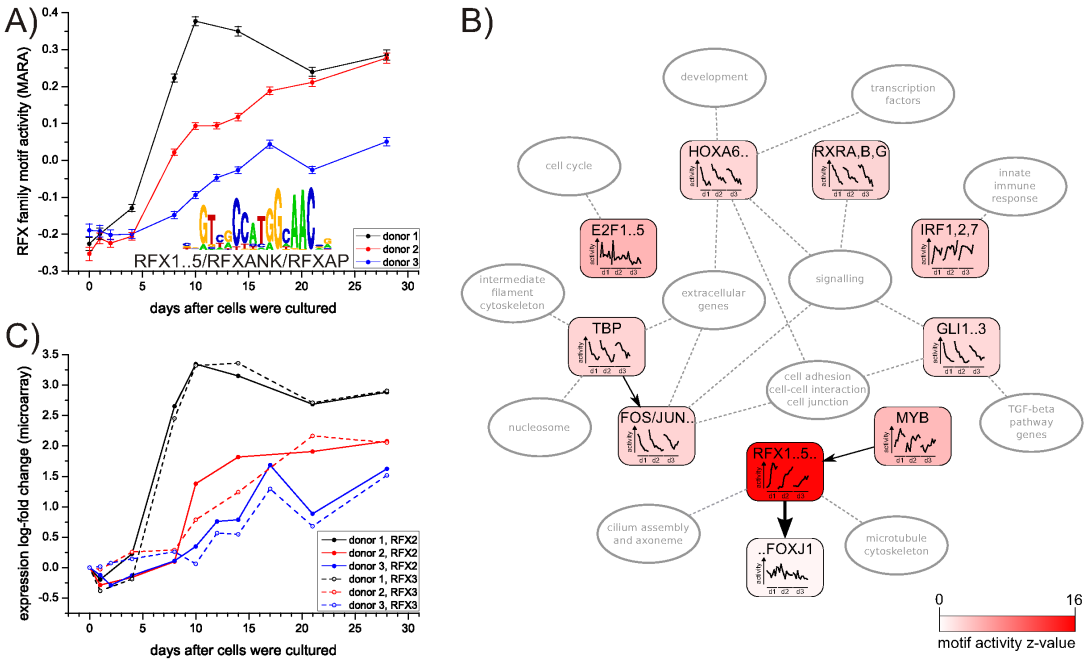


Figure 4.4: Mucociliary differentiation **A**: Inferred RFX motif activity profile in mucociliary differentiation in bronchial epithelial cells from three independent donors (black, red, and blue lines). **B**: Key predicted regulators and their targets in the mucociliary differentiation. Selected top motifs are shown together with predicted interactions between them and pathways/functional categories that are enriched among predicted targets of these motifs. The intensity of the color corresponds to the z-score of the motif, its time-dependent activity for each donor is indicated inside the node, and thickness of the edges corresponds to the target score S_{pm}. **C**: mRNA expression profiles of the RFX2 (solid) and RFX3 (dashed) genes across the differentiation (colors of the donors as in panel **A**).

RFX3 activates the FOXJ1 TF during this process. Interestingly, ISMARA also predicts that RFX directly upregulates FOXJ1 in this system. An interesting novel prediction is that RFX2 is directly regulated by the TF MYB (Fig. 4.4B). This prediction is supported by the observation that the RFX2 promoter is known to contain Myb sites and is directly regulated by A-myb in spermatogenesis [60].

As indicated in Fig. 4.4B, ISMARA additionally predicts that, in this system, IRF1,2,7 upregulates innate immune response genes, and that a short spike of E2F activity up-regulates cell-cycle genes at day 1. Finally, there is a group of motifs (TBP, FOS_FOS{B,L1}_JUN{B,D}, RXR{A,B,G}, HOX{A6,A7,B6,B7}, and GLI1..3) whose targets are progressively down-regulated across the differentiation time course. The targets of these motifs are generally enriched for extracellular proteins involved in cell adhesion, cell-cell junctions, and signaling. More specifically, targets of GLI1..3 involve genes from the TGFβ pathway, targets of TBP involve nucleosomal and intermediate filament cytoskeletal genes, and targets of the homeodomain motif (HOX{A6,A7,B6,B7}) are enriched for developmental genes and transcription factors. The genes in these pathways are most likely involved in the transition of the tissue from squamous to columnar epithelial that occurs during this differentiation. Thus, in contrast to the methods used in the original study [57], ISMARA predicts which regulators are directing various aspects of this differentiation including ciliogenesis, the innate immune response, and the transition from squamous to stratified epithelial.

4.2.5 Epithelial-Mesenchymal Transition: including microRNAs in core regulatory networks

To illustrate ISMARA's ability to integrate the role of both TFs and miRNAs in the gene regulatory network, we took advantage of data from a system in which miRNAs are known to play important regulatory roles: the epithelial-to-mesenchymal transition (EMT). Recently, mRNA expression measurements were performed in duplicate on epithelial and 3 independently-isolated mesenchymal subpopulations within immortalized mammary epithelial cells [61]. After running ISMARA on this data (results at http://ismara.unibas.ch/supp/dataset5/ismara_report/), we used replicate-averaging to identify regulators that most consistently and significantly explain the mRNA expression differences between epithelial and mesenchymal cells (results at http://ismara.unibas.ch/supp/dataset5/averaged_report/).

Interestingly, much of what is known about EMT (reviewed by Polyak and Weinberg [62]) is again captured by ISMARA's results. Among the top regulators that ISMARA infers in this system are SNAI1..3, ZEB1, and a family of miRNAs consisting of hsa-miR-141 and hsa-miR-200a (sharing the same seed sequence), that have been shown to form a regulatory network essential for EMT. The predicted activity changes of these regulators are consistent with the extant literature. Namely, the decrease in SNAI1..3 and ZEB1 activity (which indicates a reduced level of their predicted targets) in mesenchymal subpopulations is consistent with the fact that both of them are mainly acting as repressors and are transcriptionally up-regulated in the mesenchymal state. The miR-141 and miR-200a miRNAs are known to be down-regulated in the mesenchymal state, causing the mRNA levels of their targets to increase, which is consistent with the positive change in activity predicted by ISMARA. Known regulatory interactions between these factors are also uncovered by ISMARA. For instance, ZEB1 is the top predicted target of the miR-141/200a miRNAs and existing literature confirms that the direct regulation of ZEB1 by miR-200 is critical in EMT [63–65]. Similarly, E-cadherin (or CDH1) is the 3rd and 4th top target gene of the ZEB1 and SNAI1..3 motifs, respectively, and indeed this gene is an epithelial marker known to be targeted by both SNAIL transcription factors [66] and by ZEB1 [67]. These key predictions by ISMARA are summarized in Fig. 4.5.

The activity of the family containing the hsa-miR-125a/b and hsa-miR-4319 miRNAs is the most significantly reduced miRNA family in EMT. This suggests that these miRNAs play a role in mesenchymal cells, consistent with observations that miR-125b promotes invasive tumor characteristics [68].

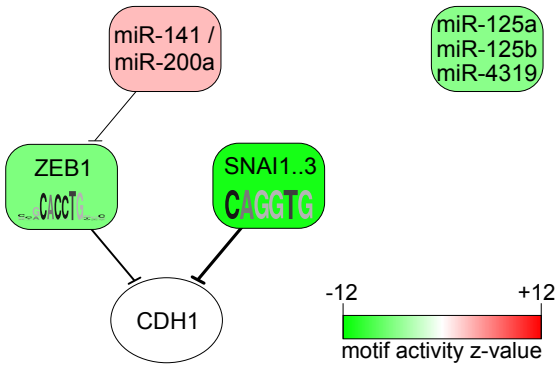


Figure 4.5: Core TF and miRNA regulatory interactions in the epithelial-to-mesenchymal transition, as predicted by ISMARA. Each rectangular node corresponds to a regulatory motif with its color indicating the significance of the change in activity when going from the epithelial to mesenchymal state (z-value defined as $z = (A_{m,\mathrm{mes}} - A_{m,\mathrm{epi}})/\sqrt{\delta A^2_{m,\mathrm{mes}} + \delta A^2_{m,\mathrm{epi}}}$). Green/Red indicates targets of the motif are down/up-regulated in the mesenchymal state. Both Zeb1 and Snail are predicted to target the E-cadherin (CDH1) promoter. Note that all interactions shown are repressive.

4.2.6 TF activities effecting chromatin state: analysis of ChIP-seq data

Beyond analyzing gene expression data, motif activity response analysis can be applied to modeling any signal along the genome in terms of the local occurrence of TFBSs. Indeed, in a recent work [69] we applied the MARA approach to ChIP-seq data mapping the dynamics of tri-methylation at lysine 27 of histone 3 (H3K27me3) and identified TFs involved in recruiting this epigenetic mark set by the Polycomb system. In ISMARA the analysis of ChIP-seq data has now been completely automated. In particular, given a ChIP-seq data set, ISMARA quantifies the signal at all promoters across all samples and models this in terms of the TFBSs at each promoter. For the details of ISMARA's processing and normalization of the ChIP-seq data we refer to the Supplementary Methods.

To illustrate ISMARA's results on ChIP-seq data, we make use of data from the ENCODE project in which, besides gene expression, 9 different chromatin modifications were measured across 8 different cell types [70] (all modifications and cell types are listed in Suppl. Tables 4 and 5). We first ran ISMARA separately on each of the 10 data sets, i.e. expression and 9 chromatin modifications (see Suppl. Table 6 for the URLs of the results on all data sets). Exploring these results we observed that motifs that are highly significant for explaining differences in levels of a particular chromatin mark across tissues, were often also highly significant for explaining *mRNA expression* differences. This was particularly the case for methylation of lysine 4 on histone H3 (H3K4me2, H3K4me3), for acetylation of histone H3 (H3K9ac, H3K27ac), and for tri-methylation of lysine 36 on histone H3 (H3K36me3). For example, Fig. 4.6A shows the activity profiles for these marks for the SNAI motif, which is recognized by the Snail TFs. Other examples of activity profiles of motifs with high significance for these marks are shown in Suppl. Fig. 17. As these figures show, for each motif, the activity profile for expression is highly similar to those of all of these histone marks. Indeed, it has been well recognized that these chromatin marks are associated with promoter activity [71], and several recent studies have shown that the levels of these marks can be used to predict gene expression levels [72–74].

To investigate the correlations between the levels of the different chromatin marks more quantitatively, we performed principal component analysis (PCA) of the distribution of the 10 different marks across all promoters, separately for each sample (Suppl. Methods). Strikingly, we find that in each sample, the first PCA component explains the majority of the variance across promoters, typically explaining around 60% of the total variance (Suppl. Fig. 18). Moreover, we find that the first PCA component looks virtual identical

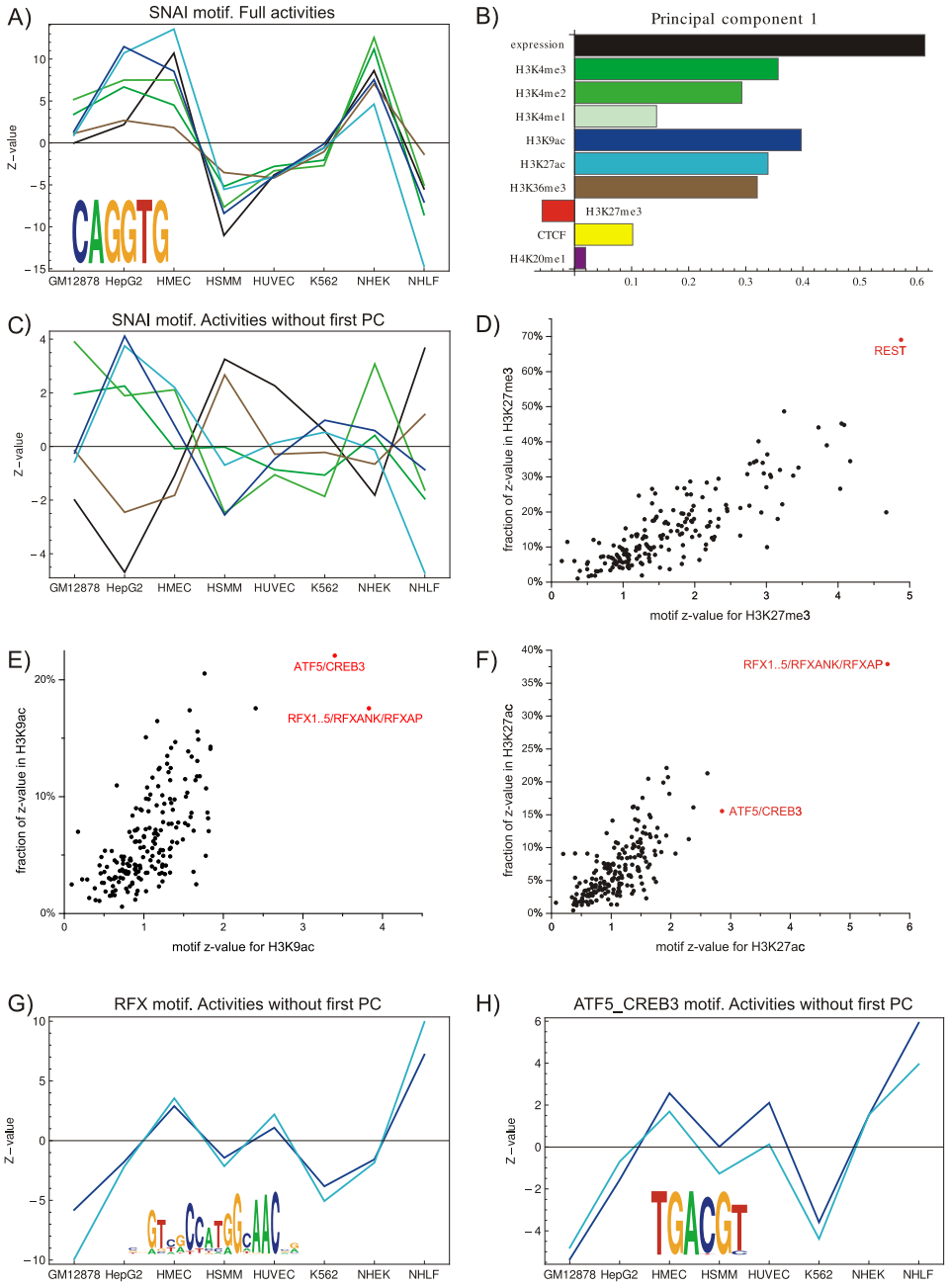


Figure 4.6: ISMARA predicts TFs involved in recruiting specific chromatin marks. **A**: Activity across cell types of the Snail motif for explaining expression (black), and levels of the chromatin marks H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown). **B**: First principal component explaining the majority of variation in chromatin mark levels across all cell types. The bars indicate the relative contributions to the principal component of each mark. **C**: Motif activities of the Snail motif, as in panel A, but after removal of the first principal component. **D**: Z-values and specificities (see text) of motifs for explaining H3K27me3 levels. The REST motif, with both highest z-value and highest specificity, is indicate in red. **E**: As in panel D, for H3K9ac levels. The two most significant motifs are shown in red. **F**: As in panels D and E, for H3K27ac levels. **G**: Activity, after removal of the first principal component, of the RFX motif for explaining H3K9ac (dark blue) and H3K27ac (light blue) levels. **H**: As in panel G, for the ATF5_CREB motif.

for each sample (Suppl. fig. 18) and Fig. 4.6B shows the first principal component obtained using PCA on the pooled data from all cell types. These findings strongly suggest that there is a single variable which corresponds roughly to 'promoter activity', which captures a large fraction of the variation in all chromatin mark levels at the promoter. In addition, the fact that this first principal vector is identical in all tissues suggests that the relative levels of the different marks in this first principal vectors result not from tissue-specific but from general factors, e.g. conceivably they may result from the general transcription machinery recruiting chromatin modifying enzymes.

The first principal vector has its highest positive component along the expression axis showing that, as expected, the expression level of the gene is most strongly aligned with its 'promoter activity'. The known activation-associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3, also all have a strong positive component in the 'promoter activity' vector. The H3K4me1 mark, which has recently been identified as a mark associated with enhancers when not accompanied by H3K4me3 [75], has a weaker positive component in the 'promoter activity' vector, as does the level of binding of the CTCF transcription factor, which is generally associated with open chromatin [76]. The known repressive mark H3K27me3, which is set by the developmentally important Polycomb system [77], indeed has a negative component in the promoter activity vector. Finally, the H4K20me1 mark shows little contribution to the first PCA component.

In summary, the PCA analysis has shown that there is a single vector in the 10-dimensional space of expression and chromatin marks that represents the general activity of a promoter and captures almost two-thirds of the variation in the levels of all marks across promoters. As a consequence, whenever a given motif contributes significantly to explaining mRNA expression in a sample, it will also contribute significantly to general promoter activity, and thereby to many of the chromatin marks. This also explains why the activity profiles of motifs that significantly explains expression are all highly correlated (Fig. 4.6A and Suppl. Fig. 17). Thus, the effect of general promoter activity on all chromatin marks confounds identification of TFs that are involved in affecting specific marks, and it would thus be beneficial to remove it. To this end we separated the activity of each motif into a part along the first PCA component, i.e. the one associated with general promoter activity, and the remaining parts along all other components. We then discarded the part of the activity along the first PCA component. As illustrated in Fig. 4.6C and Suppl. Fig. 17, after removal of the first principal component, there are no longer any obvious correlations in the remaining motif activity profiles for different activating marks. We then analyzed the remaining motif activities to identify motifs that contribute to the levels of a particular chromatin mark, independent of the motif's effect on general promoter activity.

For each motif and each mark, we calculated z-values for the remaining activity with respect to each chromatin mark. In addition, we quantified, for each motif and each chromatin mark, a 'specificity' which measures the fraction of its overall significance that is associated with the mark (Suppl. Methods). Strikingly, we find that for many of the marks, the motifs that most significantly affect the mark are also among the most specific for that mark. For example, REST is the motif with the highest z-value for H3K27me3 levels, and is also by far most specific for H3K27me3 (Fig.4.6D). Indeed, in recent work [69] we showed that REST is involved in recruiting this mark during the differentiation of murine embryonic stem cells into pyramidal neurons, specifically at the neural progenitor state. With respect to the two acetylation marks, i.e. H3K9ac and H3K27a, we find that the same two motifs, i.e. RFX and ATF/CREB, are most significant for both these marks (Fig. 4.6E and F). It is well known that ATF/CREB TFs can recruit histone acetylases (HATs) such as CREB binding protein (CBP) and p300 [78], and for RFX TFs it has also been established that they can recruit HATs at particular promoters [79]. Our results thus suggest that recruitment of HATs by TFs bound to ATF/CREB and RFX motifs make an important contribution to genome-wide histone acetylation. Moreover, the activity profiles of these motifs for H3K9ac and H3K27ac are highly similar, suggesting that these two marks may be recruited through a common or highly overlapping pathways. Supplementary Fig. 19 shows the most significant motifs for each of the other marks. Among the additional predictions made by ISMARA is that the PITX motif is associated with both mono- and di-methylation of lysine 4 of histone 3. This prediction is supported by recent biochemical evidence that PITX2 can recruit methyltransferases that

methylate H3K4 [80]. As expected, CTCF is the most significant motif explaining CTCF binding. ISMARA also makes several predictions that are completely novel, as far as we have been able to determine: It predicts that the hepatocyte nuclear factors HNF1A and HNF4A have the most significant effect on the levels of the H3K36me3 mark, which is known to be set by elongating RNA polymerase [81, 82], and that YY1 and NF-Y most significantly explain variations in H4K20me1 levels.

4.3 Discussion

Just how crucial gene regulatory circuits are in animals is evident when we remind ourselves that every cell in a multi-cellular organism has essentially the same genome, and that the phenotypic differences between cell types largely reflect differences in gene expression. The eventual goal of computational modeling of gene regulatory networks is to have realistic models of the physico-chemical interactions involved on a genome-wide scale, that accurately predict observed expression dynamics. For some very well-characterized systems of moderate size, such explicit biophysical models now appear within reach. For example, for the early anterio-posterior body patterning in *Drosophila* relatively realistic models are able to roughly capture the spatial expression patterns of dozens of cis-regulatory modules in terms of the concentration profiles of $5-10$ TFs [83, 84].

However, for the vast majority of systems our knowledge is far too rudimentary to make such detailed modeling viable. For example, an exciting recent development is the ability to reprogram cells from one differentiated state into either a stem cell state [85] or another differentiated state [86], by over-expressing or silencing specific regulatory factors. Although factors that can trigger the reprogramming cascade are known and increasing amounts of high-throughput data are available for these systems, very little is known about the regulatory networks that ultimately control these differentiation processes. The question faced by a computational biologist when analyzing such systems is how to make progress in identifying the key gene regulatory interactions given little specific knowledge of the system, and the enormous number of components potentially contributing to the system.

The advent of high-throughput technologies now allows the routine measurement of genome-wide mRNA expression across conditions, and such data in principle provide the opportunity to systematically investigate gene regulation on a genome-wide scale. Such investigations require sophisticated computational approaches and, not surprisingly, a vast literature of methods has emerged for analyzing such genome-wide expression data, ranging from explicit regulatory network models to 'black box' machine learning methods that mainly aim to capture abstract patterns in the data. Within the computational systems biology community it is sometimes implicitly assumed that the purpose of computational models of gene regulatory networks is to accurately predict gene expression patterns [87]. However, for most systems our current knowledge is far too rudimentary to expect that explicit regulatory network models can successfully model genome-wide expression patterns. Moreover, in predicting gene expression, realistic regulatory network models are often outcompeted by *ad hoc* machine learning approaches [88]. However, such approaches provide little or no insight into the underlying regulatory networks. In our opinion, the challenge is not so much to develop models that fit gene expression patterns most accurately, but to develop methods that can exploit high-throughput data to gain new insights into the underlying regulatory processes. To achieve this, the computational methods should help guide subsequent experimental efforts by prioritizing which regulatory factors are likely key players in the system, and making concrete predictions of the regulatory interactions they engage in, i.e. predictions that are directly amenable to experimental follow-up.

The Integrated System for Motif Activity Response Analysis (ISMARA), that we have presented here provides such a computational approach. Using only gene expression or chromatin state (ChIP-seq) data as input, ISMARA makes concrete predictions regarding key regulators and their regulatory interactions. Moreover, in contrast to many computational methods which require dedicated computational experts to apply, ISMARA is completely automated and provides its results in a user-friendly web-interface. In this way, ISMARA em-

powers experimentalists to infer concrete hypotheses about the genome-wide regulatory interactions acting in their system of interest, and use these to help guide more detailed experimental investigations.

That motif activity response analysis is a powerful method for reconstructing regulatory interactions from high-throughput information was already demonstrated in its original application, i.e. the reconstruction of the core regulatory network of a differentiating human cell line [13]. More recently the same approach was applied in several collaborations [89–93], showing that, in every case, ISMARA successfully inferred key regulators and their regulatory interactions *ab initio*. The applications in this work not only further confirm that, in systems where key regulatory interactions are already known, ISMARA successfully infers them, but also provide a large collection of novel regulatory predictions across different systems in human and mouse, e.g. novel regulators that are disregulated in cancers, novel regulatory interactions in the inflammatory response, master regulators of the mucociliary differentiation, etcetera. Moreover, the applications highlight several of the advantages of ISMARA. First, the fact that ISMARA infers a regulator's activity from the behavior of its targets means that non-transcriptional activity changes, i.e due to post-translational modifications, changes in cellular localization, or interactions with co-factors, can also be readily detected. Second, when a regulator's activity is transcriptionally regulated, this can help identify the relevant TF, e.g. as we did in the mucociliary differentiation by identifying RFX2 and RFX3 from the family of RFX TFs, and it can also indicate whether a regulator is acting as a repressor or an activator.

Beyond providing sorted lists of targets of each motif, the Gene Ontology analysis and the automatic visualization of the STRING network of target genes is typically very helpful in identifying the biological functions and pathways that are targeted by a particular regulator in a particular system. The links, for each predicted target, to the individual binding sites on the genome [6] provide precise predictions of the DNA segments through which the regulatory interactions are implemented, allowing for targeted validation experiments. Similarly, ISMARA's predictions of direct regulatory interactions between the key regulatory motifs provide concrete hypotheses regarding the regulatory circuitry that is acting in a given system, e.g. the predicted regulatory feedbacks between NFκB, IRF, and PRDM1, or the prediction that MYB is an upstream activator of RFX in the mucociliary differentiation.

Apart from the fact that miRNAs form an important separate regulatory layer in gene expression regulation, there are many indications that the actions of miRNAs and TFs are tightly integrated and interlinked [94–96]. By integrating both TF and miRNA regulation within an automated computational inference procedure, this allows researchers to generate hypotheses regarding the interplay of miRNAs and TFs for their system of interest. ISMARA's successful inference of the key regulatory interactions between miRNAs and TFs in EMT demonstrate its capability in this regard.

Finally, it has become clear that, especially in higher eukaryotes, regulation of gene expression involves a tight interplay and feed-back between the actions of TFs and chromatin state, with chromatin state affecting accessibility of the TFs to their sites, and TF binding being an important mechanism for recruiting chromatin modifiers that locally alter the chromatin state. In a recent work [69] we demonstrated that motif activity response analysis, applied to ChIP-seq data measuring histone modifications, can successfully identify key TFs that are involved in dynamic regulation of chromatin state. Here we have applied ISMARA to chromatin state data from the ENCODE project and provided novel predictions for, among other things, regulatory factors involved in recruiting histone acetylations.

There are of course several limitations to our approach. First, we follow Bussemaker and others [15, 16] in using a simple linear model to relate predicted TFBSs to expression patterns. The main advantage of this approach is that the model is very robust, e.g. not sensitive to wrongly predicted TFBSs or to the noise in the microarray and sequencing data. In addition, in contrast to most non-linear models, the linear model can be exactly solved, even for very large numbers of promoters and samples, so that we are guaranteed to have identified the optimal solutions. However, it is clear that it would be desirable to include saturation of the gene expression response to changes in TF activity. A second limitation is that MARA assumes that a given TF acts either mainly as an activator or mainly as a repressor whereas it is clear that some TFs can act as an activator on some targets and as a repressor on other. Indeed, it has been recently shown [97] that allowing

such dual function of TFs can significantly increase correlation coefficients between model predictions and measurement. Explicitly considering higher order constellations of TFBSs, e.g the occurrence of pairs or triplets of TFBSs for particular combinations of TFs, is another obvious extension that we are currently evaluating.

In mammals, sequence-specificities are available for only about 350 of the about 1500 TFs. Thus, it is clear that an important direction for improvement would be to obtain more comprehensive data on the sequence specificity of TFs. Recent developments in protein array technology [98], and the dramatic decrease in cost of ChIP-seq experiments make it highly likely that significant amounts of such data will become available over the coming years and we plan to use these to expand the set of regulatory motifs in ISMARA on a regular basis. In addition, whenever ChIP-seq data are available for a particular TF in a particular system, these can be used to in place of the TFBS predictions to identify target promoters directly. Indeed, we successfully used this approach in our analysis of REST's role in Polycomb recruitment [69].

ISMARA currently focuses solely on predicted TFBSs in proximal promoters, ignoring the effects of distal enhancers. The main reason for this that, in contrast to promoters, accurate genome-wide maps of distal enhancers have not been available. However, the recent realization that active enhancers exhibit characteristic chromatin modification patterns [99], DNA methylation patterns [100], and more generally DNA accessibility patterns [101], has paved the way for accurate, genome-wide identification of distal enhancers. Once a set of relevant enhancers has been identified, it is straight forward to predict TFBSs within these enhancers and incorporate these into the model.

One of the ultimate goals is to understand how regulatory interactions determine the dynamics of gene expression, and how stable 'attractors' corresponding to individual cell types are established. The direct regulatory interactions between motifs that ISMARA predicts provide a first indication of interactions that may be crucial for the observed regulatory dynamics. A key challenge in the coming years is to go beyond analysis at individual time points and develop causal models of the regulatory networks controlling the dynamics of gene expression.

4.4 Methods

Although most of the individual steps in the computational analysis employed in ISMARA are conceptually straight forward, the quality of the final results depends on many details in the computational 'protocols', and we have invested large efforts into optimizing these. For space considerations, we provide all detailed methods in the Supplementary methods and here only summarize the key steps.

ISMARA relies on our annotation of promoteromes in human and mouse, which we obtained using a combination of deepCAGE data [12] and known transcripts. For each promoter, we extracted 500 bps upstream and downstream of the TSS, and orthologous segments in 6 other mammals. The 7 orthologous sequences were then multiply aligned using T-Coffee [102]. We curated a collection of ≈ 200 WMs representing ≈ 350 mammalian TFs using data from the JASPAR [4] and TRANSFAC [5] databases, additional motifs from the literature, and our own analysis of ChIP-chip and ChIP-seq data. Binding sites were predicted using the MotEvo algorithm [8] and were performed separately for CpG and non-CpG promoters. In addition, we estimated a prior probability profile as a function of position relative to TSS for each motif. For miRNA targeting, we used the predictions of TargetScan [9] of target sites in the 3' UTR sequences of transcripts. Using the association between transcripts and promoters the miRNA target sites were associated with promoters. The miRNA predictions encompassed ≈ 100 conserved miRNA seed families. The end result of these regulator-target predictions was a site-count matrix $\mathbf{N}$, with elements N_{pm} corresponding to the estimated total number of functional binding sites for motif m in promoter p (where motifs include both TF WMs and miRNA seed families). Raw microarray and RNA-seq data were processed using standard normalization procedures. To estimate an expression profile E_{ps} for each promoter p, we use collections of known transcripts from human and mouse. We associate each promoter with all known transcripts starting at

or very near the promoter, and intersect RNA-seq reads and microarray probes with the transcripts. ChIP-seq reads are directly intersected with promoter regions, extended to the length of 1000 bp both upstream and downstream of the TSS.

To fit motif activities A_{ms} using equation (4.1) we assume that the deviations between the model and the measurements E_{ps} are Gaussian distributed with unknown variance σ^2 in each sample. To avoid overfitting we use a Gaussian prior over motif activities A_{ms} and set the variance of this prior so as to maximize generalization accuracy in a cross-validation test. Using SVD to obtain the multi-variate Gaussian posterior of motif activities, we obtain both estimated motif activities A^*_{ms} and associated (marginal) error bars δA_{ms}. The significance of each motif m is summarized by a z-like statistic:

$$z_m = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{A^*_{ms}}{\delta A_{ms}}\right)^2}, \tag{4.2}$$

where S is the number of samples. To average motif activities over subsets of samples (for example replicates) we use a Bayesian procedure that estimates both the mean activity across a subset of samples, as well as its variation. Using these estimates new error bars δA_{ms} and motif z-scores z_m are calculated.

To predict the targets of a motif m we measure, for each promoter p with predicted binding sites for m, the decrease of the quality of the fit upon removal of motif m from the model, and quantify this by a log-likelihood ratio S_{pm}. Finally, enrichment of targets within particular Gene Ontology categories is done by selecting all targets where inclusion of motif m substantially helps predicting the expression levels ($S_{pm} > 1$) and performing a standard hypergeometric test. Target networks between motifs are constructed by drawing a link from motif m to m' whenever m is predicted to target one of the promoters associated with a TF that is associated with motif m'.

4.4.1 Materials

The publicly available data sets of gene expression profiling were obtained from Gene Expression Omnibus: time course of HUVEC after TNF treatment (GSE9055), mucociliary differentiation of airway epithelial cells (GSE5264), Novartis (GNF) SymAtlas (GSE10246, GSE1133), epithelial and mesenchymal subpopulations within immortalized human mammary epithelial cells (GSE28681), ENCODE ChIP-seq (GSE26386) and expression profiling (GSE26312) in human cell lines. Microarray files from the NCI-60 were downloaded from the project web page (http://genome-www.stanford.edu/nci60/).

4.5 Supplementary Methods

4.5.1 Human and mouse promoteromes

The central entities whose regulation is modeled by ISMARA are *promoters*. When analyzing expression data, be they micro-array or RNA-seq, ISMARA estimates and models the expression profiles of individual promoters, and when analyzing ChIP-seq data ISMARA models the chromatin state of genomic regions centered on promoters. Thus, the first step in the analysis consists of the construction of reference sets of promoters in human and mouse. To make a comprehensive list of promoters we used two sources of data: deepCAGE data, i.e. next-generation sequencing data of 5' ends of mRNAs [103, 104], and the 5' ends of all known mRNAs listed in GenBank.

Using CAGE data from a considerable set of human and mouse tissues, we recently constructed genome-wide human and mouse 'promoterome' [12] consisting of a hierarchy of individual transcription start sites (TSSs), transcription start clusters (TSCs) of nearby co-regulated TSSs, and transcription start regions (TSRs), which correspond to clusters of TSCs with overlapping proximal promoter regions. As the basis of our promoter sets we started with the sets of TSCs, i.e. local clusters of TSSs whose expression profiles are proportional to each other to within experimental noise, as identified by deep-CAGE.

As the currently available CAGE data do not yet cover all cell types in human and mouse, a substantial number of cell type-specific promoters are not represented within this set of TSCs. We thus supplemented the TSCs with all 5' ends of mRNAs, using the BLAT [105] mappings from UCSC genome browser web site [106]. To avoid transcripts whose 5' ends are badly mapped, we filtered out those for which more than 25 bases at the 5' end of the transcript were unaligned. We then produced reference promoter sets by iteratively clustering the TSCs with the 5' ends of mRNAs as follows: Initially each TSC and each 5' end of an mRNA forms a separate cluster. At each iteration the pair of nearest clusters are clustered, with the constraint that there can be at most one TSC per cluster. That is, we never cluster two TSCs together as our previous analysis in [12] has already established that each TSC is independently regulated. This iteration is repeated until the distance between the closest pair of clusters is larger than 150 base pairs. Note that we thus chose the length of sequence wrapped by a single nucleosome, i.e. roughly 150 base pairs, as an *ad hoc* cut-off length for two TSSs to belong to a common promoter. The reasoning behind this choice of cut-off, is that, on the one hand, we have empirically observed that co-expressed TSSs can spread over roughly this length-scale and, on the other hand, that it is not implausible that the ejection of a single nucleosome near the TSS may be responsible for setting this length scale. In any case, the resulting promoters are not sensitive to the precise setting of this cut-off (data not shown). Finally, inspection of the results showed, especially in ubiquitously expressed genes, many apparent TSSs from Genbank that appear downstream of both the TSSs identified by deep-CAGE and the annotated RefSeq transcripts. It is highly likely that many of these apparent TSSs are due to cDNA sequences that were not full length. Indeed, only a small fraction of the transcripts in the database of mRNAs underwent expert curation, and truncated transcripts are likely common. To avoid such spurious apparent TSSs we removed all clusters which did not contain at least one curated transcript (RefSeq) or a TSC. Finally, since a sequence of at least one associated transcript is necessary to estimate a promoter's expression level from either RNA-seq or micro-array data, we also discarded all promoters that consisted solely of a TSC.

For human, the resulting reference promoter set had $36'383$ promoters, of which $13'265$ contained both a TSC and at least one RefSeq transcript, $14'538$ contained only a TSC together with non-RefSeq transcripts, and $8'580$ had at least one RefSeq transcript and potentially non-RefSeq transcripts, but no TSC. For the mouse genome, the corresponding numbers are: $34'050$ promoters in total, $8'578$ RefSeq-only, $12'303$ TSC-only, and $13'169$ with both a TSC and at least one RefSeq transcript. These reference promoters sets cover almost all known protein-coding genes in human and mouse.

Finally, as we discussed in [12], mammalian promoters clearly fall into two classes associated with high and low content of CpG dinucleotides, and these promoter classes have clearly distinct architectures, i.e.

different lengths, different numbers of TSSs per promoters, and different distributions of transcript factor binding sites (TFBSs). We classified all promoters into a high-CpG and low-CpG class based on both the CG content and the CpG content in the proximal promoter, as described in [12]. In the TFBS prediction below we perform separate predictions for high-CpG and low-CpG promoters.

4.5.2 A curated set of regulatory motifs

We use standard position dependent weight matrices (WMs) to represent regulatory motifs, i.e. the sequence specificities of TFs. Each WM is named for the TFs that are annotated to bind its site. For some motifs the names correspond to multiple TFs which are all assumed to bind to the same sites. We used a partly manual curation procedure whose details were first described in [13]. For completeness, we here also give a description of this curation procedure.

For a number of reasons regarding data quality and annotation ambiguities, the construction of a set of position-specific weight matrices (WMs) for mammalian transcription factors is rife with problems that, in our opinion, do not currently have a clean solution. Therefore, our procedures necessarily involve several subjective choices, judgments, and hand-curation, which are certainly far from satisfactory.

Our main objectives were

1. To remove redundancy, we aim to have no more than 1 WM representing any given TF. Whenever multiple TFs have WMs that are statistically indistinguishable or when their DNA binding domains are virtually identical, then we use only one WM for that set of TFs.

2. To associate WMs with TFs based on the sequences of their DNA binding domains. That is, we obtain lists of TFs that can plausibly bind to the sites of a given WM by comparison of DNA binding domain sequences of TFs known to bind to the sites with those of all other TFs.

3. Re-estimation of WMs using genome-wide predictions of regulatory sites in the proximal promoters of CAGE TSSs.

The input data for our WM construction consisted of

1. The collection of JASPAR vertebrate WMs plus, for each WM, the amino acid sequence of the TF that JASPAR associates with the WM [107].

2. The collection of TRANSFAC vertebrate WMs (version 9.4) and the amino acid sequences of all vertebrate TFs in TRANSFAC that are associated with those WMs [4].

3. A list of 1322 human TFs (Entrez gene IDs) and their amino acid sequences (from RefSeq).

4. A list of 483 Pfam IDs corresponding to DNA binding domains and their Pfam profiles [108].

We decided not to include 6 TRANSFAC motifs, which were constructed out of less than 8 sites: M00326 (PAX1, PAX9), M00619 (ALX4), M00632 (GATA4), M00634 (GCM1, GCM2), M00630 (FOXM1), M00672 (TEF). TRANSFAC often associates multiple WMs with a single human TF. Although there undoubtedly are cases where a single TF can have multiple distinct modes of binding DNA, and could therefore be realistically represented by multiple WMs, we believe that for the very large majority of TFs it is more realistic to describe the DNA binding specificity of the TF with a single WM. Indeed, a manual inspection of cases in which TRANSFAC associated multiple WMs with a single TF shows that these WMs are typically highly similar and appear redundant. Therefore, we decided to remove this redundancy and for each TF with multiple WMs in TRANSFAC we choose only a single 'best' WM based on TRANSFAC's own matrix quality annotation, or WM information score when there were multiple WMs with the same quality score. The information score of a WM is given by 2 times the length of the WM minus its entropy in bits.

We next aimed to obtain, for each human TF, a list of WMs from JASPAR/TRANSFAC, that can potentially be associated to this TF. To do this we aim to find, for each TF, which motifs from JASPAR/TRANSFAC are associated with a TF that has a highly similar DNA binding domain. To this end we we ran Hmmer [109] with the DNA binding domain (DBD) profiles from Pfam to extract the DBDs from all TFs (E-value cut-off 10^{-9}) associated with either JASPAR or TRANSFAC matrices. We then represented each such TF with the union of its DNA binding domain sequences. Next we used BLAT to map the DBDs of all TFs associated with JASPAR/TRANSFAC matrices against the entire protein sequences of all human TFs. For each human TF we then extracted a list of all JASPAR/TRANSFAC matrices for which the DBDs of at least one associated TF has a significant BLAT hit (default parameters) against the TF sequence. For each human TF the associated WMs were ordered by the percent identity of the hit, i.e. the fraction of all amino acids in the DBDs that map to matching amino acids in the TF.

From this data we created a list of 'necessary WMs' as follows. For each human TF we obtain the JASPAR WM with the highest percent identity in the DBDs of an associated TF. If there is a TRANSFAC WM with a higher percent identity than any JASPAR TF we record this WM as well. Thus, the necessary WMs are those that are the best match for at least one human TF. This list yielded 381 WMs representing 980 human TFs (often the same WM is the best match for multiple TFs). Manual inspection indicated that a lot of redundancy (essentially identical looking WMs) remained in this list. First we often have both a TRANSFAC and a JASPAR WM for the same TF and moreover often there are multiple TFs, each with its own WM, that look essentially identical. We thus want to fuse WMs in the following situations

1. Different WMs for TFs with identical or near identical DBDs.

2. WMs that are statistically indistinguishable, predict highly overlapping sets of sites, and are associated with TFs that have similar DBDs.

For each pair of WMs we obtained three similarity measurements

1. The percent identity of the DBDs of the TFs associated with the WMs. If there are multiple TFs associated with a WM we take the maximum over all TF pairs.

2. The overlap of the binding sites predicted by each WM. We use MotEvo to predict TFBSs in all proximal promoters and we calculate what fraction of predicted TFBS positions are shared between the two WMs.

3. A statistical measure of the similarity of the two WMs. Here we take the two sets of sites that define the two WMs and calculate the likelihood-ratio of these sites assuming they either derive from a single underlying WM or assuming that the set of sites for each WM derives from an independent WM.

For each of these three criteria we set a cut-off: 95% identity of the DBDs, 60% overlap of predicted TFBSs, and a likelihood-ratio of e^{40}. Using single-linkage clustering, we cluster all WMs whose similarity is over the cut-off for at least 1 of these three criteria. The resulting clusters were then all checked manually and whenever the linkage was dubious we split the cluster. That is, we took a conservative attitude towards removing redundancy and only kept clusters when we were convinced the WMs were essentially identical. For each cluster we then constructed a new WM by aligning the WMs in the cluster so as to optimize the information content of the resulting fused WM, which is obtained by simply summing the counts across each column in the alignment.

Finally, we used MotEvo [10] to predict TFBSs for all WMs in the multiple-species alignments of all human proximal promoters. We then constructed new WMs from the list of predicted TFBSs for each WM, weighing each predicted site with its posterior probability (which incorporates position-specific prior probabilities, as described below). The number of top-scoring sites was chosen manually for each motif and was between 100 and 4000 sites, in most cases being 200 or 500 sites.

At this point we excluded one TRANSFAC motif M00395 (HOXA3, HOXB3, HOXD3) which had very low information content and predicted predicted only very low-probability sites. We additionally excluded the motifs M00480 (TOPORS) and M00987 (FOXP1), which were unrealistically specific and (in case of M00987) predicted stretches of poly(T).

For a few TFs we obtained more recent WMs from the literature (SP1, OCT4, NANOG, SOX2, XBP1, PRDM1, and the RXRG dimer) and we used these to replace the corresponding WM in the list.

We improved several motifs by running MotEvo on TF ChIP-seq data: SRF, STAT1/3, REST and ELK1/4/GABPA/GABPB1. Some other motifs were obtained by predicting *de novo* using the Phylogibbs algorithm [110] on ChIP-seq data: SPI1, CTCF, OCT4, SOX2 and NANOG.

For a few motifs JASPAR has recently updated or introduce new motifs which were based on high-throughput data and we included these motifs.This is the case for FOXA2, KLF4, EWSR1-FLI1, FEV, NR4A2. We also removed MA0118, as it had been discarded in JASPAR data base.

Our final list contains 189 WMs. For each final WM there is an ordered list of associated human TFs, ordered by percent identity of the DBDs of TFs known to bind sites of the WM and the DBDs of the TF. We then checked this list of associations by hand and for each WM cut-off the list of associated human TFs manually. In total 340 human TFs are associated with our 189 WMs. The corresponding mouse orthologous TFs were selected using the MGI data base [111]. The entire set of WMs and mapping to associated TFs is available from the SwissRegulon website (http://www.swissregulon.unibas.ch).

4.5.3 Transcription factor binding site predictions

After creating reference promoter sets and curating a set of mammalian regulatory motifs we next predicted TFBSs in the proximal promoter regions of each promoter. Analysis of sequence conservation in the neighborhood of TSSs (see [12]) and experimentation with TFBS prediction in regions of different lengths around TSSs indicated that a reasonable balance between sensitivity (i.e. including relevant binding sites) and specificity (avoiding too many false positive predictions) can be obtained by predicting TFBSs in a 1 kilobase region around the TSSs of each promoter.

For each promoter, we thus extended the promoter sequence spanned by its cluster of TSSs by 500 bp upstream and 500 bp downstream. We denote this as the *proximal promoter region* of a promoter. We then extracted the sequence of the reference species, i.e. human or mouse and orthologous regions from 6 other mammals (human or mouse, rhesus macaque, cow, dog, horse, and opossum) using pairwise BLASTZ [112] alignments. For each promoter, we multiply aligned the orthologous regions using T-Coffee [102].

To obtain a phylogenetic tree for these mammalian species, with branch lengths corresponding to the expected number of substitutions per neutrally evolving site, we used methods described previously [113]. Briefly, we first obtained the topology of the tree from the UCSC genome browser [114]. Then, for each pair of species we made pairwise alignments of the coding regions of orthologous genes and extracted all third positions in fourfold-degenerate codons of amino acids that are conserved between the two species. Using these fourfold-degenerate positions we estimated a pairwise distance for each pair of species. Finally, we estimated the lengths of the branches in the phylogenetic tree as those that minimize the square-deviations between the implied pairwise distances and the pairwise distances estimated from the fourfold-degenerate positions. The resulting tree structure is shown in Suppl. Fig. 4.7.

The multiple sequence alignments were then used together with the phylogenetic tree and the collection of WMs as an input for TFBSs predictions using the MotEvo algorithm [10]. Given a multiple alignment, MotEvo considers all ways in which the sequence of the reference species can be segmented into 'background' positions, 'binding sites' for one of the supplied WMs, and 'unknown functional elements' (UFEs). The likelihood of alignment columns assigned to background are calculated under a model of neutral evolution along the specified phylogenetic tree. The likelihood of alignment segments assigned to be a site for a given WM are calculated by first estimating which of the species have retained a site for the WM (based on the WM scores of the individual sequences) and then applying an evolutionary model in which substitution

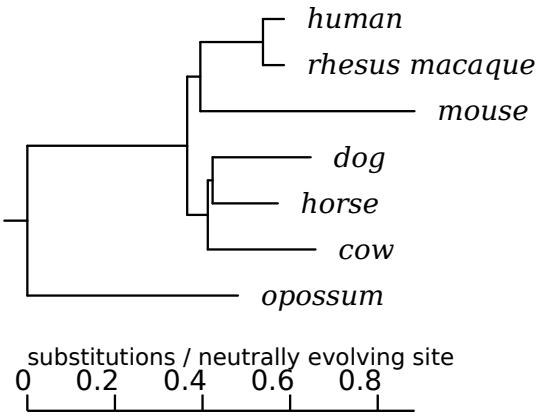


Figure 4.7: The phylogenetic tree used by MotEvo for the transcription factor binding site predictions that are used by ISMARA.

rates are set so as to match the sequence preferences of the WM. Finally, segments assigned to be UFEs are assumed to evolve under *unknown* purifying selection constraints on the sequence, which is implemented by treating them as sites for an unknown WM, which is treated as a nuisance parameter that is integrated out of the likelihood. Finally, MotEvo assigns, at each position of the alignment and for each WM, a posterior probability that a site for the corresponding WM occurs at this position.

Since most motifs show clear positional preferences relative to TSS, we implemented a position-dependent binding prior probability distribution for each motif which we fitted by maximum likelihood using expectation-maximization. Since high-CpG and low-CpG promoters have highly distinct configurations of TFBSs, we estimated the position-dependent prior probability distributions separately for high-CpG and low-CpG promoters.

The final result of this analysis is a matrix $\mathbf{N}$, with N_{pm} the total number of predicted sites for motif m in promoter p, i.e. the sum of the posterior probabilities of the individual sites. To reduce the probability of spurious predictions, we set $N_{pm} = 0$ whenever the sum of the posteriors of all sites was less than 0.1.

4.5.4 Associating miRNA target sites with each promoter

Apart from incorporating the effects of TFBSs in promoters, ISMARA also integrates the effects of miRNAs in its modeling of expression levels. To this end, we needed to obtain a set of predicted miRNA target sites for each promoter. We base our predictions on the miRNA target predictions of TargetScan using preferential conservation scoring (aggregate P_{CT}) [9] which has shown consistently high performance in various benchmark tests. As opposed to focusing on individual miRNAs, TargetScan groups miRNAs that have identical subsequences at positions 2 through 8 of the miRNA, i.e. the 2-7 seed region plus the 8th nucleotide, and provides predictions for each such seed motif. We will treat these seed motifs exactly like the regulatory motifs (WMs) for TFs, i.e. a miRNA seed motif can be associated with multiple miRNAs. TargetScan provides predictions for 86 mammalian miRNA seed motifs in total.

TargetScan P_{CT} provides a score for each seed motif and each RefSeq transcript. To obtain a 'site count' N_{pm} for the number of sites of miRNA seed motif m associated with promoter p we average the TargetScan P_{CT} scores of all RefSeq transcripts associated with the promoter p. Finally, the miRNA seed motif site counts N_{pm} are simply added as columns to the site count matrix $\mathbf{N}$ with site counts of TFBSs.

4.5.5 Expression data processing

When using expression data from oligonucleotide microarrays, the raw probe intensities are corrected for background and unspecific binding using the Bioconductor packages affy [115], oligo [116], and gcrma [117], depending on the type of the particular microarray used. The micro-arrays that are currently supported by ISMARA are listed in supplementary table 4.1.

Microarray	Organism	Producer
HG-U133A	*Homo sapiens*	Affymetrix
HG-U133B	*Homo sapiens*	Affymetrix
HG-U133_Plus_2	*Homo sapiens*	Affymetrix
HG-U133A_2	*Homo sapiens*	Affymetrix
HuGene-1_0-st-v1	*Homo sapiens*	Affymetrix
HuGene-1_1-st-v1	*Homo sapiens*	Affymetrix
HT_HG-U133A	*Homo sapiens*	Affymetrix
HT_HG-U133B	*Homo sapiens*	Affymetrix
HT_HG-U133_Plus_PM	*Homo sapiens*	Affymetrix
Mouse430_2	*Mus musculus*	Affymetrix
Mouse430A_2	*Mus musculus*	Affymetrix
MOE430A	*Mus musculus*	Affymetrix
MOE430B	*Mus musculus*	Affymetrix
MoGene-1_0-st-v1	*Mus musculus*	Affymetrix
MoGene-1_1-st-v1	*Mus musculus*	Affymetrix
HT_MG-430A	*Mus musculus*	Affymetrix
HT_MG-430B	*Mus musculus*	Affymetrix
MG_U74Av2	*Mus musculus*	Affymetrix
MG_U74Bv2	*Mus musculus*	Affymetrix
MG_U74Cv2	*Mus musculus*	Affymetrix

Table 4.1: Microarrays currently supported by ISMARA

For its further analysis, ISMARA uses the logarithms of the probe intensities. For a given sample, the histogram of log-intensities is generally bimodal, with the modes corresponding to probes of non-expressed and expressed genes. The probes are classified as expressed or non-expressed in each sample separately by fitting a two-component Gaussian mixture model to the log-intensity data using the Mclust R package [118, 119]. Probes that are consistently non-expressed are filtered out from further processing; a probe is considered not to be expressed if in *all* the samples the probability of it belonging to the expressed class is below 0.4. Subsequently, the intensity values are quantile normalized across all input samples.

Micro-array probes can hybridize to multiple transcripts, belonging to different genes, or different isoforms of one gene, and we decided not to rely on transcript annotations of a micro-array producer. Instead, we comprehensively mapped the probe sequences to the set of all transcripts that are associated with our reference set of promoters. Note that we thus also ignore the annotation of probes into probe sets. To calculate the expression of a promoter we average the log-expression levels of all probes that map to one (or more) of the transcripts associated with the promoter (i.e. the start of the transcript is a member of the cluster of starts that defines the promoter). The expression level of the promoter is then a weighted average of the expression levels of these probes, where a probe that maps to n different transcripts obtains a weight $1/n$. That is, in general, a probe can map to multiple transcripts.

When ISMARA uses RNA-seq for input expression data, it expects the RNA-seq data to be provided as genome alignments of the reads to the hg19 or mm9 genome assembles in BED format. The loci of the

mapped reads are then intersected with the genome alignments of all transcripts that are associated with reference promoters. A read is associated with a particular transcript if it falls entirely into any of its exons. We thus unfortunately discard a fraction of reads which originated from exon-exon junctions. However, the alternative of using read mappings to transcripts would require the user to map to the exact same set of transcripts as used by ISMARA and this is impractical. In the future ISMARA may be extended to include mapping of raw reads.

To obtain an expression level for each promoter ISMARA calculates a weighted average over all reads mapping to the transcripts associated with the promoter. The weighting results from multiple mappings at two levels. Firstly, a single read can map to multiple genomic loci and, secondly, a single locus may intersect multiple transcripts that are associated with multiple promoters. When a read maps to n genomic loci, we assign a weight of $1/n$ to each locus. If that locus intersects transcripts of m different promoters, then this reads contributes a final weight of $1/(nm)$ to the expression of each promoter. For a given promoter p and sample s, the total weight w_{ps} is the sum of the weights of all the reads that intersect one of the transcripts associated with promoter p. The expression E_{ps} of promoter p in sample s is then given by

$$E_{ps} = \log\left[\frac{w_{ps}}{N_s}\right], \tag{4.3}$$

where N_s is the total number of reads in sample s, which map to any of the transcripts associated with a reference promoter. Note that this weighting procedure is robust to redundancy in the transcript sets. For example, when a promoter is associated with k highly overlapping transcripts, then a read mapping within the exons of these transcripts will get assigned to all these transcripts, with a weight $1/k$ for each. When the total weight w_{ps} of the promoter is calculated, these k are then summed back and will in the end contribute precisely 1 read. Note also that because ISMARA models promoter expression *changes* across conditions rather than absolute levels, there is no need to account for differences in transcript lengths (i.e. these just cause a shift in log-space which cancels out when considering expression changes).

4.5.6 ChIP-seq data processing

Apart from modeling expression dynamics, ISMARA can also process ChIP-seq data to automatically model chromatin state (or TF binding) changes at promoters genome-wide. Examples of such chromatin state data include histone occupancy, histone modifications, TF binding and DNAse1 hypersensitivity in promoter regions. After several experiments, we found that integrating the chromatin signal from a region of 2000 bps centered on the TSS of each promoter gives the most robust results. To obtain a chromatin state level E_{ps} of promoter p in sample s, we calculate the sum r_{ps} of the reads that map entirely within this region around promoter p and transform to the log-space after adding a pseudocount:

$$E_{ps} = \log_2\left(r_{ps} + \frac{N_s l}{L}\right), \tag{4.4}$$

where the second term is a pseudo-count, N_s is the total number of reads mapped to the genome in sample s (the number of lines in the BED file), $l = 2000$ is the length of the regions, and L is the total length of the genome. Note that this pseudo-count is precisely the number of reads that would be expected if all N_s reads where distributed uniformly over the genome. We set to pseudo-count to this value to make the pseudo-count roughly of the same size as the read-count from background reads in regions where the chromatin mark in question does not appear. The rational is that, in regions where there are only background reads, statistical fluctuations may cause the read-counts r_{ps} to change significantly from sample to sample. By adding a constant pseudo-count of roughly the same size these fluctuations are effectively dampened. More formally, this pseudo-count results within a Bayesian context if we use a Dirichlet prior with an expected density l/L for each region.

4.5.7 Motif activity fitting.

We model log-expression (or ChIP-seq signal) value E_{ps} of a promoter p in sample s as a linear function of the site-counts N_{pm} for all motifs m associated with the promoter, i.e. either TFBSs in the proximal promoter region or miRNA binding sites in the 3' UTRs of associated transcripts. In each sample s, the contribution of the sites N_{pm} to E_{ps} is given by the (unknown) *motif activity* A_{ms}. That is, we fit a model of the form:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \tag{4.5}$$

where $\tilde{c}_s$ and c_p are sample and promoter-dependent constants, and we assume that the noise is Gaussian distributed with an unknown variance σ^2 that is the same for all promoters and in all samples. Under these assumptions we find the following expression for the likelihood of the expression data given the site-counts, motif activities and sample and promoter-dependent constants: We first maximize this expression with respect to all the constants c_p and $\tilde{c}_s$, and substitute these with their *maximum likelihood* estimates. After doing this we obtain:

$$P\left(E \;\middle|\; A^{'}, N, \sigma\right) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{ps}\left(E^{'}_{ps} - \sum_m N^{'}_{pm} A^{'}_{ms}\right)^2}{2\sigma^2}\right], \tag{4.6}$$

where P is the number of promoters, S is the number of samples, the $N^{'}_{pm}$ are a motif-normalized site-counts $N^{'}_{pm} = N_{pm} - \langle N_m \rangle$, with $\langle N_m \rangle$ the average site-count per promoter for motif m, the A'_{ms} are sample-normalized activities $A^{'}_{ms} = A_{ms} - \langle A_m \rangle$, i.e. with $\langle A_m \rangle$ the average activity of motif m across the samples, and the $E^{'}_{ps}$ are sample- and promoter-normalized expression values $E^{'}_{ps} = E_{ps} - \langle E_p \rangle - \langle E_s \rangle + \langle\langle E \rangle\rangle$. That is the log-expression matrix $E^{'}_{ps}$ is normalized such that all its rows and columns sum to zero, the activities $A^{'}_{ms}$ are normalized such that the average activity over all samples is zero, i.e. $\sum_s A^{'}_{ms} = 0$, and the site-counts $N^{'}_{pm}$ are normalized such that the average count over all promoters is zero, i.e. $\sum_p N^{'}_{pm} = 0$.

To avoid over-fitting we assign a symmetric Gaussian prior to each motif activity, i.e. the joint prior for all activities is given by:

$$P\left(A^{'} \;\middle|\; \lambda, \sigma\right) \propto \prod_{ps} \exp\left[-\frac{\lambda^2}{2\sigma^2} \sum_m A^{'2}_{ms}\right], \tag{4.7}$$

where the constant λ^2 sets the width of prior distribution relative to the width of the likelihood function. Using this prior with the likelihood derived above, the posterior distribution of motif activities takes the form:

$$P\left(A^{'} \;\middle|\; E, N, \sigma, \tau\right) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{p,s}\left(\left(E^{'}_{ps} - \sum_m N^{'}_{pm} A^{'}_{ms}\right)^2 + \lambda^2 \sum_m A^{'2}_{ms}\right)}{2\sigma^2}\right]. \tag{4.8}$$

Since equation (4.8) factorizes into independent expressions for the different samples, it is enough to consider one sample at a time. The posterior distribution for the motif activities in a particular sample takes the general form of a multi-variate Gaussian centered around $A^{'*}_{ms}$:

$$P\left(A^{'}_s \;\middle|\; E, N, \sigma\right) \propto \sigma^{-P} \exp\left[-\frac{\sum_{m\tilde{m}}\left(A^{'}_{ms} - A^{'*}_{ms}\right) W_{m\tilde{m}} \left(A^{'}_{\tilde{m}s} - A^{'*}_{\tilde{m}s}\right) + \chi^2_s}{2\sigma^2}\right], \tag{4.9}$$

where the χ^2_s is the unexplained part of variance in sample s

$$\chi^2_s = \sum_p \left(E^{'}_{ps} - \sum_m N^{'}_{pm} A^{'*}_{ms}\right)^2, \tag{4.10}$$

and the matrix W is given by

$$W_{m\tilde{m}} = \sum_p \left(N'_{pm} N'_{p\tilde{m}} + \lambda^2 \delta_{m\tilde{m}} \right). \tag{4.11}$$

Finally, the *maximum a posteriori* (MAP) estimates A'^{*}_{ms} can be found by minimizing the expression in the numerator of equation (4.8) using standard numerical procedures for ridge regression. ISMARA performs this calculation by singular value decomposition of the N' matrix.

Setting λ through cross-validation

Both the MAP estimates A'^{*}_{ms}, and the matrix $W_{m\tilde{m}}$ are functions of λ. The constant λ^2 represents the ratio between the a priori expected variance of activities, to the average squared-deviation of the model from the expression data (which results from both error in the model, noise in the expression measurements, and biological noise). In general λ will depend on the measurement platform used, i.e. microarray, RNA-seq, or ChIP-seq, and also on the samples used, because the true variance in motif activities will depend on the variance in the E_{ps} across the samples. Thus, the appropriate value of λ will generally not be known in advance and ISMARA therefore includes a method for automatically setting λ from the data. To determine the optimal λ ISMARA uses a $80/20$ cross-validation scheme. The set of promoters is divided randomly into two sets, with one containing 80% of all promoters (the 'training set') and the other the remaining 20% (the 'test set'). The training set of promoters is used for fitting the motif activities while the quality of the fit is evaluated on the test set. ISMARA then finds the value of λ that minimizes the average squared-deviation of the expression levels in the test set from those predicted by the model. We denote this optimal value of λ by λ^*.

Error bars on motif activities

Apart from the MAP estimates A'^{*}_{ms} ISMARA also determines the uncertainties associated with these estimates. Since σ in Eq. 4.9 is not known, we integrate it out with a suitable scale-invariant prior $P(\sigma) \propto \frac{1}{\sigma}$.

$$\begin{aligned} P\left(A'_s \,\middle|\, E, N, \lambda\right) &= \int_{\sigma=0}^{\infty} P\left(A'_s \,\middle|\, E, N, \sigma, \lambda\right) P(\sigma) \mathbf{d}\sigma \\ &\propto \frac{\Gamma\left(\frac{P}{2}\right)}{\left[\sum_{m\tilde{m}} \left(A'_{ms} - A'^{*}_{ms}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s} - A'^{*}_{\tilde{m}s}\right) + \chi_s^2\right]^{\frac{P}{2}}} \\ &\propto \exp\left[- \frac{P \sum_{m\tilde{m}} \left(A'_{ms} - A'^{*}_{ms}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s} - A'^{*}_{\tilde{m}s}\right)}{2\chi_s^2} \right], \end{aligned} \tag{4.12}$$

where the last proportionality is a very good approximation when the number of promoters is large. Note that this is again a multi-variate Gaussian distribution. The covariance matrix of this Gaussian posterior distribution is given by:

$$C_{m\tilde{m};s} = \frac{\left(W^{-1}\right)_{m\tilde{m}} \chi_s^2}{P} \tag{4.13}$$

As is well known, given this multi-variation Gaussian form, the marginal distribution for a single motif activity A'_{ms} will be Gaussian distributed with standard-deviations $\delta A'_{ms}$ given by the square root of the corresponding diagonal term of the covariance matrix, i.e.

$$\delta A'_{ms} = \sqrt{C_{mm;s}} \tag{4.14}$$

We define the overall *significance* of a motif m as the average squared ratio between fitted activities and their standard deviations (z-values)

$$z_m = \sqrt{\frac{1}{S}\sum_s \left(\frac{A'^*_{ms}}{\delta A'_{ms}}\right)^2}. \tag{4.15}$$

4.5.8 Processing of replicates

Careful studies typically involve experimental replicates to account for the part of variability in the readout which is not under direct experimental control. ISMARA allows users to indicate which samples correspond to replicates and will automatically calculate averaged motif activities and error bars across these replicates. To perform this analysis the user should first upload all samples and perform the standard analysis. On the results page ISMARA provides a link to a page where users can interactively annotate which samples are replicates. In addition, if the replicates came in clearly defined batches, for example, when a time-course was performed multiple times, then the user can also indicate this. Once all samples are annotated ISMARA can then perform motif activity averaging across the replicates. Note that this approach can easily be extended beyond replicates, i.e. the user can arbitrarily divide the samples into groups and ISMARA will automatically calculate average motif activities and associated standard-deviations for each group of samples.

Here we describe how activities within a group are averaged. For a given group G of samples and a particular motif, we assume that its activities A_s in samples $s \in G$ are given by a mean activity $\bar{A}^g$ plus some deviation δ_s, i.e

$$A_s = \bar{A}^g + \delta_s, \tag{4.16}$$

where we assume that the prior probability of δ_s is Gaussian distributed with (unknown) standard-deviation σ_g, i.e

$$P(\delta_s|\sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g}\exp\left[-\frac{1}{2}\frac{\delta_s^2}{\sigma_g^2}\right]. \tag{4.17}$$

Thus, given the mean activity $\bar{A}^g$ in the group, the probability to have activity A_s in a particular sample s from the group is

$$P(A_s|\bar{A}^g,\sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g}\exp\left[-\frac{1}{2}\frac{\left(A_s-\bar{A}^g\right)^2}{\sigma_g^2}\right]. \tag{4.18}$$

Using the input data, ISMARA has inferred the motif activity A_s to have expected value A_s^* with standard-error δA_s for each sample s. That is, once the dependence on all other activities is integrated out, the probability of the expression data D conditioned on the motif activity A_s is a Gaussian with standard-deviation δA_s, i.e.

$$P(D|A_s) = \frac{1}{\sqrt{2\pi}\delta A_s}\exp\left[-\frac{1}{2}\frac{(A_s-A_s^*)^2}{(\delta A_s)^2}\right]. \tag{4.19}$$

Using the expressions for $P(D|A_s)$ and $P(A_s|\bar{A}^g,\sigma_g)$ we can calculate the probability of the data D given the mean activity $\bar{A}^g$ and standard-deviation σ_t by integrating over all unknown A_s:

$$P(D|\bar{A}^g,\sigma_g) = \prod_{s\in G}\left[\int_{-\infty}^{\infty} P(D|A_s)P(A_s|\bar{A}^g,\sigma_g)\mathbf{d}A_s\right]. \tag{4.20}$$

These integrals can be performed analytically and we obtain

$$P(D|\bar{A}^g,\sigma_g) = \prod_{s\in G}\frac{1}{\sqrt{2\pi(\sigma_g^2+\sigma_s^2)}}\exp\left[-\frac{(A_s^*-\bar{A}^g)^2}{2(\sigma_g^2+\sigma_s^2)}\right]. \tag{4.21}$$

Although, formally, we should integrate this expression over the unknown standard-deviation σ_g as well, this integral unfortunately cannot be performed analytically. Therefore, we estimate the integral simply by finding the value σ_g^* that maximizes $P(D|\bar{A}^g, \sigma_g)$. Assuming a uniform prior for the mean activity $\bar{A}^g$ of the samples in the group, we then finally obtain an expression for the posterior probability $P(\bar{A}^g|D)$ which we characterize by its mean $\langle \bar{A}^g \rangle$ and standard-deviation $\delta \bar{A}^g$. That is, $\langle \bar{A}^g \rangle$ is the inferred average motif activity for the samples within the group, and $\delta \bar{A}^g$ is the error-bar on this average activity. This mean and error-bar of the activity for the 'group' of samples are given by

$$\langle \bar{A}^g \rangle = \frac{\sum_{s \in G} \frac{A_s^*}{(\sigma_g^*)^2 + \sigma_s^2}}{\sum_{s \in G} \frac{1}{(\sigma_g^*)^2 + \sigma_s^2}}, \tag{4.22}$$

and

$$\delta \bar{A}^g = \sqrt{\frac{1}{\sum_{s \in G} \frac{1}{(\sigma_g^*)^2 + \sigma_s^2}}}. \tag{4.23}$$

Finally, we assign significances z_m to each motif completely analogously as before, but now averaging over all groups, i.e.

$$z_m = \sqrt{\frac{1}{|G|} \sum_g \left(\frac{\langle \bar{A}^g \rangle}{\delta \bar{A}^g} \right)^2}, \tag{4.24}$$

where $|G|$ is the number of groups. A motif will have a high significance z_m when its motif activities in each group vary little relative to their mean in the group, and are large relative to the original error-bars.

4.5.9 Target predictions

In order to infer motif activities A_{ms}, ISMARA assumes that all promoters with predicted target sites for a motif m will respond to changes in motif activity, i.e. in proportion to the predicted number of sites N_{pm}. This is a reasonable assumption when inferring motif activities, as the activities A_{ms} depend on the statistics of all promoters with sites for motif m. However, in a given condition or system, it is likely that only a subset of the promoters with sites for a motif m are in fact regulated by this regulator. This might be due to a limited accessibility, dependence of particular co-factors, weaker affinity of a site, etcetera. Thus, when we aim to predict individual target promoters of a given motif m, we not only use the binding site predictions N_{pm}, but also evaluate at which promoters the activities A_{ms} contribute to explaining the profiles E_{ps}.

To quantify if a given promoter p is targeted by a motif of interest m we first demand that there exists a TFBS prediction, i.e $N_{pm} > 0$. Second, we quantify the contribution of m to the fit of the expression/chromatin state profile E_{ps}. The most rigorous approach to quantifying the effect of motif m on promoter p is to calculate both the probability of the entire data set, i.e. the profiles E_{ps} across all promoters and samples, with the original site-count matrix $\mathbf{N}$, and a site-count matrix $\tilde{\mathbf{N}}$ where only the sites for motif m in promoter p are set to zero. To calculate this probability we treat all the unknown motif activities A_{ms} as well as the standard-deviation σ as nuisance parameters that are integrated out of the likelihood. That is, we formally want to calculate the ratio of probabilities

$$R_{pm} = \frac{\int_{-\infty}^{\infty} \mathbf{d}A \int_0^{\infty} \mathbf{d}\sigma P(E|N, A, \sigma)}{\int_{-\infty}^{\infty} \mathbf{d}A \int_0^{\infty} \mathbf{d}\sigma P(E|\tilde{N}, A, \sigma)}, \tag{4.25}$$

where the integrals are over all motif activities A_{ms}, and over the standard-deviations σ. Note that, when we set $N_{pm} = 0$ for promoter p and motif m, we make a very small change to the site-count matrix. That is, as there are tens of thousands of promoters and close to 200 motifs, we are changing only one of the millions of entries in the matrix. As a consequence, the inferred motif activities $A_{ms}^{'*}$ that result from the mutated matrix

$\tilde{N}$ are likely very close to those that result from the original matrix N. Similarly, the inverse covariance matrix W of the mutated matrix is likely also very close to that of the original matrix and, finally, the optimal values of the constants c_p, $\tilde{c}_s$, and the prior constant λ^* will also change very little under mutation of the matrix. To make the calculation more tractable we will make the approximation that all these quantities are *unchanged* upon mutation of the matrix. Under that approximation we have

$$P(E|A,N,\sigma,\lambda^*) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{s,m,\tilde{m}} \left(A'_{ms} - A'^*_{ms}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s} - A'^*_{\tilde{m}s}\right) + \sum_{p,s} \chi^2_{ps}}{2\sigma^2}\right], \quad (4.26)$$

where χ^2_{ps} is the squared-deviation between the observed value E'_{ps} and the predicted value, i.e.

$$\chi^2_{ps} = \left(E'_{ps} - \sum_m N'_{pm} A'^*_{ms}\right)^2 \quad (4.27)$$

And for the probability of the data with the mutated site-count matrix we have

$$P(E|A,\tilde{N},\sigma,\lambda^*) = P(E|A,N,\sigma,\lambda^*) \exp\left[-\frac{\sum_s (\chi^2_{psm} - \chi^2_{ps})}{2\sigma^2}\right], \quad (4.28)$$

where χ^2_{psm} is the squared-deviation for promoter p and sample s when motif m is removed, i.e.

$$\chi^2_{psm} = \left(E'_{ps} - \sum_{m'} \tilde{N}'_{pm'} A'^*_{m's}\right)^2 \quad (4.29)$$

In this form the integrals over the motif activities and σ can be easily performed and we find for the ratio of the probabilities

$$R_{pm} = \left(\frac{\sum_{p',s} \chi^2_{p's}}{\sum_{p',s} \chi^2_{p's} - \sum_s (\chi^2_{psm} - \chi^2_{ps})}\right)^{S(P-M)}, \quad (4.30)$$

where M is the total number of motifs. Since $P \gg M$ we approximate $P - M \approx P$ and we find approximately

$$R_{pm} = \exp\left[\frac{\sum_s (\chi^2_{psm} - \chi^2_{ps})}{\langle \chi^2 \rangle}\right], \quad (4.31)$$

where we have defined the average squared-deviation per sample/promoter combination

$$\langle \chi^2 \rangle = \frac{1}{PS} \sum_{p,s} \chi^2_{ps}, \quad (4.32)$$

and made use of the fact that $[1 - x/(SP)]^{-SP} \approx e^x$ for large SP.

In the results shown in the web-server we show, for each predicted target, the logarithm of the likelihood ratio, i.e. the score S_{pm} for motif m targeting promoter p is

$$S_{pm} = \frac{\sum_s \chi^2_{psm} - \chi^2_{ps}}{\langle \chi^2 \rangle}. \quad (4.33)$$

All targets for which this score is positive, i.e. where removing the motif from the promoter reduces the quality of the fit, are reported.

Enriched Gene Ontology categories

To analyze whether there are any Gene Ontology categories whose genes are over-represented among the targets of a motif, we use the "GO::TermFinder" Perl module [120]. The ontology files and associations between genes and categories were taken from Gene Ontology (GO) Consortium web-site [121]. As a set of target genes for motif m we include all genes associated with promoters that have a target score $S_{pm} > 0$. For microarray chips we create a background set from all the genes which are have complementary probes present on the microarray, i.e. have associated probes of the microarray according to our mappings (see Expression data processing). For RNA-seq data we take all genes associated with promoters which have mapped reads. In the web results we display all GO categories with a p-value of 0.01 or less. These p-values are corrected for multiple testing using a simple Bonferroni correction, i.e. multiplied by the number of tests performed.

4.5.10 Principal component analysis of the activities explaining chromatin mark levels

We first performed standard ISMARA analysis on the $n = 10$ data sets measuring expression and 9 different chromatin marks (ChIP-seq), across $S = 8$ cell types [70]. For each motif m, and each mark i, we thus obtained estimated activities A^i_{ms}.

We performed principal component analysis (PCA) of the expression and chromatin mark levels across all promoters, separately for each cell type. For a given sample s, let E_{pi} denote the level of mark i at promoter p (suppressing the label s for notational simplicity). We have here already column normalized these levels, i.e.

$$\sum_p E_{pi} = 0, \tag{4.34}$$

for all marks i.

Using singular value decomposition, the matrix $E = U \cdot D \cdot V^T$ can be uniquely decomposed into an orthonormal matrix U (of size $P \times n$), a diagonal positive-semidefinite matrix D (of size $n \times n$), and an orthonormal matrix V (of size $n \times n$) as:

$$E_{pi} = \sum_{k=1}^{n} U_{pk} D_{kk} V_{ik}, \tag{4.35}$$

where k denotes the index of each component, the column vectors $\vec{V}_k$ with components V_{ik} contain the principal components, and D^2_{kk} is the fraction of the variance in the E_{pi} values, i.e.

$$\mathrm{var}(E) = \frac{1}{nP} \sum_{p,i} (E_{pi})^2 , \tag{4.36}$$

that is explained by component k.

The first principal component $\vec{V}_1$, shown in Suppl. Fig. 4.24 top panels, is virtually identical in all cell types and captures approximately 60% of the collective behavior of the expression and 9 chromatin marks (8 histone modification and CTCF binding) across promoters in each sample. As discussed in the main text, this first principal component appears to capture the combination of chromatin mark levels associated with the general 'activity' of a promoter. As a consequence, the effect of a given TF on a specific chromatin mark is confounded by its effect on general promoter activity and we therefore decided to subtract it from the activity profiles of all TFs.

For the purpose of removing the first principal component from the motif activities, we will treat each motif m separately and ignore the covariances in the inferred motif activities, i.e. as we assumed previously when calculating the error bars on the motif activities in (4.14). We perform the removal one sample (cell line) at a time. A careful probabilistic analysis must be performed in order to calculate the error bars.

Let's focus on a given motif m in sample s and denote by A the vector of activities across the marks, i.e. A_i is the activity associated with mark i. In addition, let δA_i denote the standard-deviation (error-bar) of this activity. The posterior distribution $P(A|D)$ of this activity vector given the data is given by a Gaussian, i.e. as in (4.13), of the form

$$P(A|D) \propto \exp\left[-\frac{1}{2}\sum_i \frac{(A_i - A_i^*)^2}{\delta A_i^2}\right], \tag{4.37}$$

where A_i^* is the MAP estimate of the motif activity of mark i. If we introduce a diagonal matrix containing the inverse of the standard-deviation, we can write this expression in matrix-vector form:

$$P(A|D) \propto \exp\left[-\frac{1}{2}(A - A^*)^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot (A - A^*)\right], \tag{4.38}$$

where A^* is a $n \times 1$ vector of the MAP estimates and $\mathrm{diag}\left(\frac{1}{\delta A^2}\right)$ is a 10×10 diagonal precision matrix which elements are set to the inverses of motif activity variances.

Using principal components V of E (4.35) and their orthonormality $V \cdot V^T = \mathbb{1}$ this distribution can be rewritten as

$$P(A|D) \propto \exp\left[-\frac{1}{2}(A - A^*)^T \cdot V \cdot V^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot V \cdot V^T \cdot (A - A^*)\right]. \tag{4.39}$$

We can rewrite the activities in the basis of the principal vectors as $B \equiv V^T \cdot (A - A^*)$ and the precision matrix in the same basis as $M \equiv V^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot V$. In this basis the probability distribution takes the form:

$$P(B|D) \propto \exp\left[-\frac{1}{2}B^T \cdot M \cdot B\right]. \tag{4.40}$$

Note that in this basis, the inverse covariance matrix M contains off-diagonal terms.

We want to integrate out the activities along the first principal component, therefore we separate elements of B and M in the following way

$$B = \begin{pmatrix} b_1 \\ \begin{pmatrix} b_2 \\ \vdots \\ b_n \end{pmatrix} \end{pmatrix} \equiv \begin{pmatrix} b_1 \\ B_y \end{pmatrix} \tag{4.41}$$

$$M = \begin{pmatrix} m_{11} & \begin{pmatrix} m_{12} & \cdots & m_{1n} \end{pmatrix} \\ \begin{pmatrix} m_{21} \\ \vdots \\ m_{n1} \end{pmatrix} & \begin{pmatrix} m_{22} & \cdots & m_{2n} \\ \vdots & \ddots & \vdots \\ m_{n2} & \cdots & m_{nn} \end{pmatrix} \end{pmatrix} \equiv \begin{pmatrix} m_{11} & M_y^T \\ M_y & M_w \end{pmatrix}, \tag{4.42}$$

and the last equivalency holds because the matrix M is symmetric.

Using these definitions, eq. (4.40) can be expanded and rewritten to obtain:

$$\begin{aligned} P(B|D) &\propto \exp\left[-\frac{1}{2}\left(b_1^2 m_{11} + 2b_1 B_y^T \cdot M_y + B_y^T \cdot M_w \cdot B_y\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(m_{11}\left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}}\right)^2 + B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}}\right)\right] \end{aligned} \tag{4.43}$$

Where we reordered terms and completed the square to bring out that this posterior is proportional to a Gaussian with respect to b_1. It is now straightforward to integrate this probability distribution along the first

principal direction:

$$\begin{aligned} P(B_y|D) = \int_{b_1=-\infty}^{\infty} P(B|D)\mathbf{d}b_1 \quad &\propto \quad \exp\left[-\frac{1}{2}\left(B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}}\right)\right] \cdot \\ &\cdot \int_{b_1=-\infty}^{\infty} \exp\left[-\frac{1}{2} m_{11}\left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}}\right)^2\right] \mathbf{d}b_1 \qquad (4.44) \\ &\propto \quad \exp\left[-\frac{1}{2} B_y^T \cdot \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}}\right) \cdot B_y\right], \end{aligned}$$

The last proportionality holds because the Gaussian integral yields a constant (with respect to B_y). Since the covariance matrix is the inverse of the precision matrix, the covariance matrix W in the reduced $(n-1)$-dimensional space (i.e. without the first principal direction) has the form:

$$W = \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}}\right)^{-1} \qquad (4.45)$$

Finally, this covariance matrix W needs to be transformed back from the principal component basis to the original basis. To this end we use the principal components contained in columns 2 through n of the V matrix. We obtain for the final covariance matrix K in the original basis

$$K_{ij} = \sum_{k,l=2}^{n} V_{ik} W_{kl} V_{jl}. \qquad (4.46)$$

The standard deviation of activities of the i^{th} mark is given by square root of the corresponding diagonal element of this matrix

$$\delta\tilde{A}_i = \sqrt{K_{ii}}. \qquad (4.47)$$

The corrected MAP activities are obtained by first defining

$$B^* = V^T \cdot A^*, \qquad (4.48)$$

and then transforming back to the original basis using only the components along principal vectors 2 through n:

$$\tilde{A} = \sum_{k=2}^{n} V_{ik} B_k^*. \qquad (4.49)$$

The reported z-value of the i^{th} mark (we introduce back the indices for motif m and sample s omitted previously) is given by

$$z_{ms}^i = \frac{\tilde{A}_i}{\delta\tilde{A}_i} \qquad (4.50)$$

After removing the contribution of the first principal component to the motif activities, we re-calculated significance z-values z_m^i for each motif m and each mark i (x-axis in the Suppl. Fig. 4.25)

$$z_m^i = \sqrt{\frac{\sum_{s'} \left(z_{ms'}^i\right)^2}{S}}. \qquad (4.51)$$

In addition, we calculated a specificity s_m^i which measures the fraction of the overall that is associated with mark i (y-axis in the Suppl. Fig. 4.25)

$$s_m^i = \frac{z_{mk}^2}{\sum_{k'} z_{mk'}^2}. \tag{4.52}$$

That is, a motif m will be highly specific for mark i if it has a high z-value z_m^i, and low z-values for all other marks.

4.6 Fraction of variance explained by the fit

The total variance V in a data set is given by the sum of the squared normalized expression values

$$V = \frac{1}{PS} \sum_{p,s} (E'_{ps})^2. \tag{4.53}$$

After fitting the model, the average squared deviation left unexplained is given by the average of χ_{ps}^2 across all promoters and samples, i.e. as defined by equations (4.27) and (4.32). The fraction of the variance f explained by the fit is thus

$$f = 1 - \frac{\langle \chi^2 \rangle}{V}. \tag{4.54}$$

We find that the fraction of variance explained by the fit typically ranges between 4% and 14%. As an illustration, Suppl. Fig. 4.8 shows a histogram of the fraction of variance explained by the model across all samples in the GNF data set.

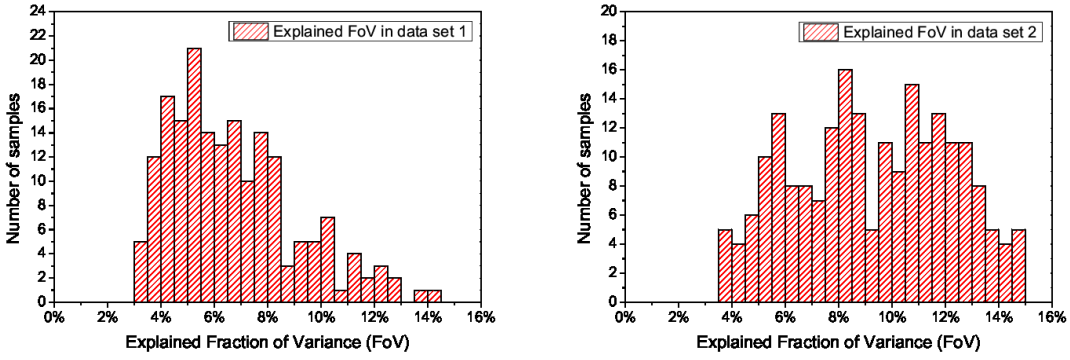


Figure 4.8: Histogram of the fraction of variance explained by the model. **Left panel:** Histogram of the fraction of variance explained for the samples in the mouse GNF atlas (data set 1). **Right panel:** Histogram of the fraction of variance explained for the samples from the human GNF atlas and the NCI-60 cancer cell lines (data set 2).

The fraction of variance explained in the samples of the second data set (human GNF atlas plus NCI-60 cell lines) is a bit larger than the fraction of variance explained in the samples of data set one (the mouse GNF atlas). It appears that this increase results from the fact that there is a relatively large (and explainable) difference in the expression profiles of the cancer cell lines and the normal cell lines.

4.7 Overview of results presented in the web-interface

To illustrate the results that ISMARA provides, we here present a number of figures, that show examples of results on the mouse GNF atlas. Note that almost all of these figures are screen shots from the actual web-interface. All the full results for the mouse GNF data are available at http://ismara.unibas.ch/supp/dataset1/ismara_report.

The main page of results that ISMARA provides for a given data set centers around a list of motifs, sorted by their significance, showing for each motif its significance, the associated TFs, a sequence logo of the motif, and a thumbnail image of its inferred activity across the samples. Supplementary Fig. 4.9 shows an excerpt from this list of motifs.

Figure 4.9: Fragment of the list of regulatory motifs sorted by their significance (z-score). The motifs are sorted from top to bottom. Shown for each motif are, from left to right, the name of the motif (which is a link to a separate page with results for the motif), its z-score, a list of associated TFs (links to NCBI pages for these genes), a thumbnail of the inferred motif activity profile, and the sequence logo of the motif.

Each motif name in this list is in fact a link to a separate page with much more extensive results for the motif. Among these more extensive results is, first of all, a figure showing the inferred motif activity (and error bars) across all samples, where the samples are ordered according from left to right, according to the user's input. Supplementary Fig. 4.10 shows the activity profile of the E2F1..5 motif across the mouse GNF samples. Note that such an ordering motif activity across samples is especially helpful when the samples come from a time course, in which case the graph shows the motif activity across time.

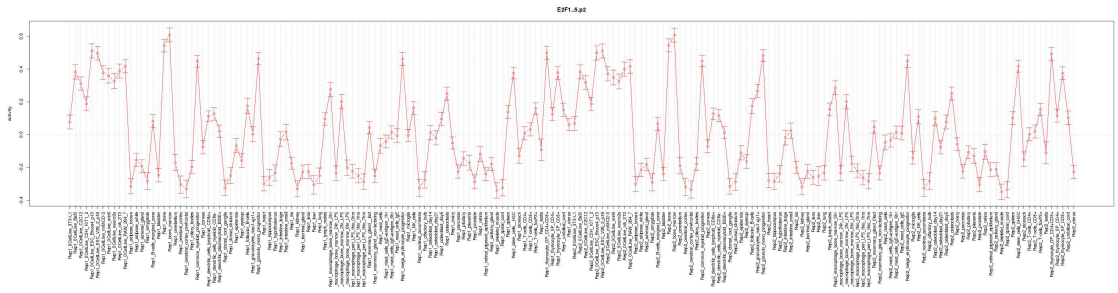


Figure 4.10: Inferred activities of the E2F1..5 motif on the mouse GNF atlas. The samples are ordered, from left to right, in lexicographic order, according to sample names input by the user. The red circles show the estimated activities A^*_{ms} and the error-bars δA_{ms} are shown as red vertical bars. Samples names are indicated on the bottom.

However, in many cases, including the GNF atlas analyzed here, there is no preferred natural ordering of the samples. In those cases it is more natural to present the motif activities with samples sorted from those in which the motif is most significantly upregulated, to those where it is most significantly downregulated. ISMARA provides such a list of motif z-values, with sample sorted from largest to smallest z-value, as shown in Suppl. Fig. 4.11 for the E2F1..5 motif. In this case, inspection of this sorted list of samples makes clear that E2F is highly upregulated in fast dividing cells, and downregulated in post-mitotic cells. The close association of E2F activity with cell proliferation is something that we have observed across many different data sets (data not shown).

The next important information provided for each motif, is a predicted list of target promoters. ISMARA provides the target promoters p for a motif m sorted by their target score S_{pm} (see section 4.5.9). As an example, the list of targets for the E2F1..5 motif is shown in Suppl. Fig. 4.12. Each row in the table corresponds to one target promoter and information shown includes the promoter ID, its score S_{pm}, associated transcripts and Entrez gene, and a description of the gene. Note that all these pieces of information are links that take the user to additional information on the promoter, the associated transcripts and gene. Note that,

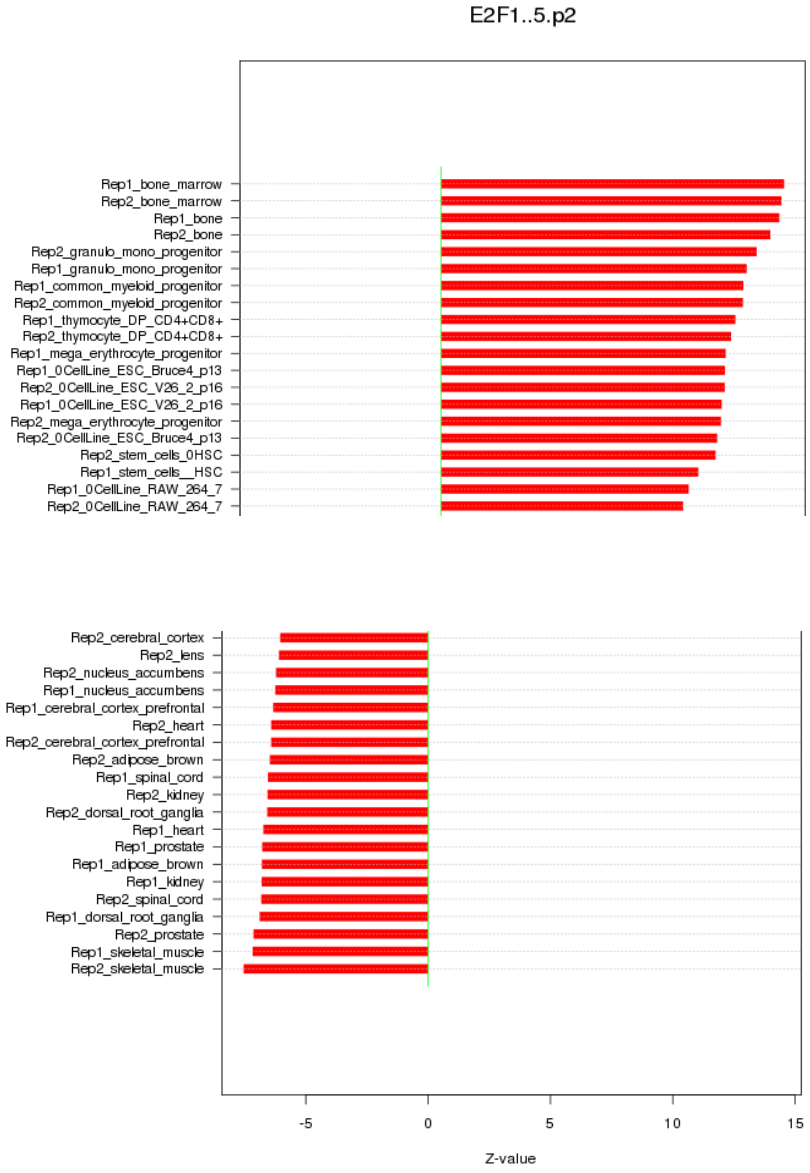


Figure 4.11: Sorted list of z-values for the E2F motif across all samples of the mouse GNF atlas. For readability, only the top 20 and bottom 20 samples are shown. Note that the samples with the highest z-values correspond to fast proliferating cells whereas the samples with the lowest z-values correspond to non-proliferating cells.

to keep the page easily viewable, by default only the top 20 targets are shown. But the user can interactively change the number of targets shown in the list. In addition, a search box allows the user to search whether a particular promoter, transcript, or gene of interest occurs within the full list of targets.

Top 20 targets:

Search:

Show 20 entries

Showing 1 to 20 of 200 entries

Promoter	Score	Refseq	Gene	Description
chr12_+_25393083	205.825	NM_009104	Rrm2	ribonucleotide reductase M2
chr10_-_68815606	190.395	NM_007659	Cdk1	cyclin-dependent kinase 1
chr1_-_130256055	178.689	NM_008567	Mcm6	minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae)
chr15_-_57966551	170.642	NM_027435	Atad2	ATPase family, AAA domain containing 2
chr8_+_77633426	170.047	NM_008566	Mcm5	minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae)
chr10_+_110182506	159.597	NM_178609	E2f7	E2F transcription factor 7
chr1_-_20810238	152.621	NM_008563	Mcm3	minichromosome maintenance deficient 3 (S. cerevisiae)
chr10_-_20880559	148.266	NM_001198914 NM_010848	Myb	myeloblastosis oncogene
chr5_-_138613028	146.050	NM_008568	Mcm7	minichromosome maintenance deficient 7 (S. cerevisiae)
chr8_+_125091889	138.459	NM_026014	Cdt1	chromatin licensing and DNA replication factor 1
chr2_+_72314235	134.555	NM_025866	Cdca7	cell division cycle associated 7
chr11_+_98769122	126.897		Cdc6	cell division cycle 6 homolog (S. cerevisiae)
chr13_-_21925343	122.491	NM_178185 NM_001177544	Hist1h2ah Hist1h2ap Hist1h2ai Hist1h2ao	histone cluster 1, H2ah histone cluster 1, H2ap histone cluster 1, H2ai histone cluster 1, H2ao
chr6_-_88848650	116.296	NM_008564	Mcm2	minichromosome maintenance deficient 2 mitotin (S. cerevisiae)
chr11_+_98769162	115.939	NM_001025779	Cdc6	cell division cycle 6 homolog (S. cerevisiae)
chr13_-_21879086	111.341	NM_178184	Hist1h2an	histone cluster 1, H2an
chr13_-_22134795	106.518	NM_178186	Hist1h2ag Hist1h2ai	histone cluster 1, H2ag histone cluster 1, H2ai
chr10_+_127669118	104.620	NM_001136082 NM_001164080 NM_001164081	Timeless	timeless homolog (Drosophila)
chr2_-_157030236	103.984	NM_001139516 NM_011249	Rbl1	retinoblastoma-like 1 (p107)
chr10_+_127669154	103.485		Timeless	timeless homolog (Drosophila)

Figure 4.12: Top target promoters of the E2F1..5 motif for the mouse GNF atlas. Targets are sorted by the log-likelihood score S_{pm}. Shown for each target promoter are the promoter ID (a link to the SwissRegulon web-browser page showing the promoter on the genome), the target score S_{pm}, associated RefSeq transcripts, associated gene symbols (links to NCBI pages), and gene names (which often provide a short description of the gene's function). By default the top 20 targets are shown but this number can be changed using the drop-down menu at the top of the table. A search box allows users to search for genes or transcripts within the entire target list.

Of particular interest is the additional information provided about each promoter, through the links with the promoter IDs. Following this link takes the user to the genome browser of our SwissRegulon database [122], showing the section containing the proximal promoter regions (500 base pairs up-stream and down-stream of the major TSS of the promoter). In this browser the user is shown all the predicted TFBSs that are used by ISMARA in its modeling of expression or ChIP-seq data. This thus allows the user to determine the precise locations of the TFBSs on the genome, through which a particular TF is predicted to target a given promoter. Supplementary Fig. 4.13 shows, as an example, the promoter of the Rrm2 gene, which is the top predicted target of the E2F1..5 motif.

Beyond a list of individual targets, a user would typically like to gain some intuition of the pathways and particular biological processes that are targeted by a particular motif. One way of visualizing the functional structure of the predicted targets of a motif, is to represent these as a network, with links between pairs of genes that are known to be functionally related. The STRING database [20] maintains a curated collection of functional links between proteins, where 'functional link' can range from direct physical interaction, to over-representation of the protein pair within abstracts of scientific articles. For any set of proteins, STRING

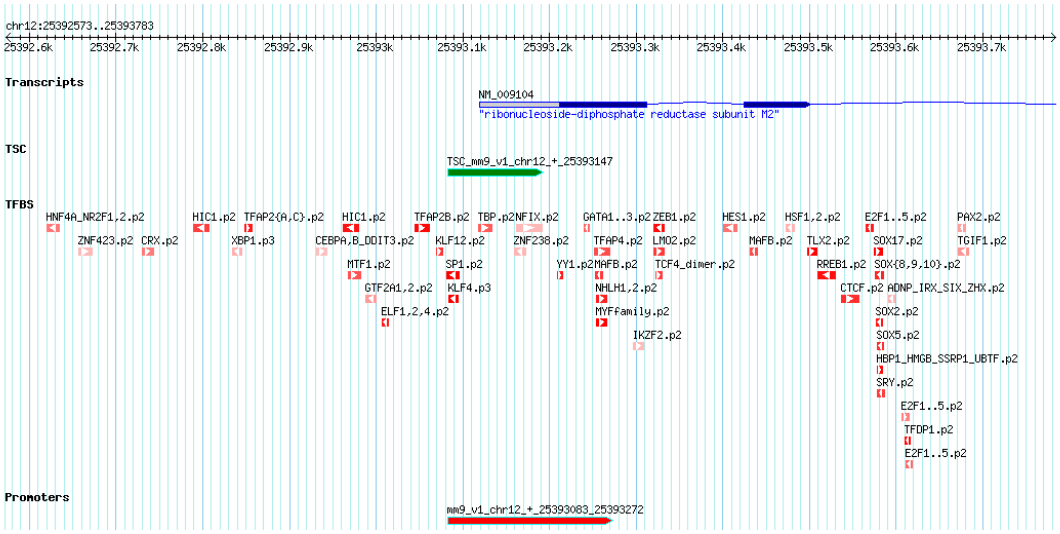


Figure 4.13: Example of a promoter region as display in the SwissRegulon genome browser. The region shown corresponds to the proximal promoter of the Rrm2 gene (the top target of the E2F1..5 motif) and this is the region that will be displayed when following the link to the promoter as displayed in Suppl. Fig. 4.12. The genome browser shows the RefSeq transcript, the promoter, the associated annotated transcript start cluster (TSC) based on the CAGE data, and all the predicted TFBSs. Here the intensity of the color indicates the posterior probability assigned to each site, and the name of the cognate motif is written above each side. The arrows inside the TFBSs indicate on which strand the motif occurs.

provides visualizations of the network of known functional interactions between these proteins, which intuitively brings out groups of proteins known to be functionally related. ISMARA provides, for each motif, such STRING network pictures of the set of predicted targets of the motif (for visibility at most the top 200 targets are shown). Supplementary Fig. 4.14 shows the STRING network for the predicted targets of E2F1..5. Note that the picture is itself a link to the STRING database, where the figure is interactive and allows the user more detailed information on each of the proteins in the network and each functional link between the proteins.

Apart from the STRING network, ISMARA also provides list of Gene Ontology categories that are enriched among the predicted targets of a motif. Lists are provided for the 'biological process', 'cellular component', and 'molecular function' hierarchies. A p-value for enrichment is calculated using a simple hypergeometric test and only categories with a p-value below 0.05 are shown. The categories are sorted by the fold-enrichment of targets relative to what would be expected by chance. As an example, Suppl. Fig. 4.15 shows the top categories of the biological process hierarchy for the E2F1..5 motif.

For many of the motifs incorporated into the ISMARA analysis, there is more than one TF that can potentially bind to sites for the motif. As a consequence, it is not always clear which individual TFs are responsible for the observed motif activity in a particular system. To help determine which TFs are most likely involved in the activity of a given motif, ISMARA provides some simple correlation analysis. In particular, a table is provided showing the Pearson correlation between the motif's activity profile and the mRNA expression profiles of each of the TFs that can bind to the sites of the motif. The TFs in the list are sorted by their p-value. Supplementary Fig. 4.16 shows the list of correlations for the E2F TFs.

For each of the correlations a link is also provided to a simple scatter plot showing the mRNA expression levels and motif activities across the samples. Supplementary Fig. 4.17 shows example scatter plots for the TFs E2F1 and E2F2, which are both significantly correlated with motif activity. The fact that both motifs correlate *positively* with motif activity strongly suggests that these TFs act as activators, i.e. as their mRNA levels go up, the expression of target genes is affected positively. To show an example of opposite behavior,

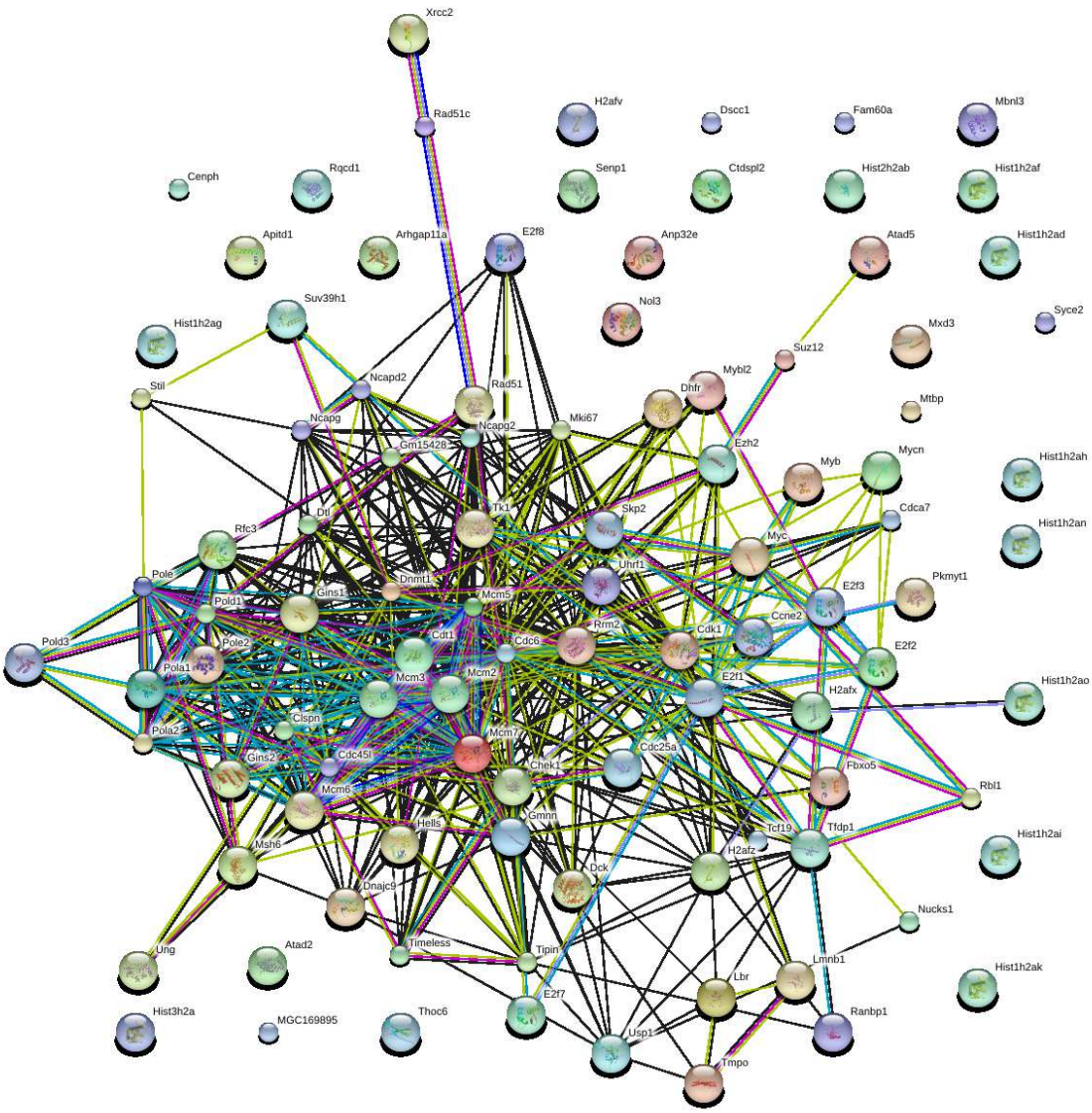


Figure 4.14: Network of target genes of the E2F motif as displayed by the STRING database [20]. Each node corresponds to a predicted target gene of the E2F1..5 motif (in the mouse GNF atlas, i.e. data set 1). Links are drawn by STRING whenever there is any evidence that the two genes may interact or be functionally linked, where evidence may range from measured direct protein-protein interaction to significant co-occurrence of the gene names within abstracts of articles.

Gene overrepresentation in process category:

Search:

Show 10 entries

Showing 1 to 20 of 160 entries

enrichment	p-value	GO term	description
18.42	1.97e-02	GO:0019985	translesion synthesis
16.57	8.61e-08	GO:0006270	DNA-dependent DNA replication initiation
15.07	4.51e-07	GO:0071897	DNA biosynthetic process
13.15	2.04e-02	GO:0000731	DNA synthesis involved in DNA repair
12.28	6.83e-05	GO:2000104	negative regulation of DNA-dependent DNA replication
11.84	1.41e-05	GO:0090329	regulation of DNA-dependent DNA replication
11.05	1.00e-02	GO:0000084	S phase of mitotic cell cycle
11.05	1.00e-02	GO:0006268	DNA unwinding involved in replication
10.05	2.10e-02	GO:0051320	S phase
9.69	2.87e-05	GO:0006297	nucleotide-excision repair, DNA gap filling
8.29	8.81e-04	GO:0007062	sister chromatid cohesion
7.37	1.18e-02	GO:0032392	DNA geometric change
7.21	2.20e-08	GO:0006261	DNA-dependent DNA replication
6.82	2.13e-29	GO:0006260	DNA replication
6.82	1.76e-03	GO:0033261	regulation of S phase
6.31	3.67e-04	GO:0008156	negative regulation of DNA replication
5.84	3.20e-04	GO:0043966	histone H3 acetylation
4.93	4.31e-04	GO:0006289	nucleotide-excision repair
4.87	1.33e-03	GO:0051053	negative regulation of DNA metabolic process
4.72	9.25e-06	GO:0006473	protein acetylation

Figure 4.15: Top over-represented categories from the Gene Ontology hierarchy of biological processes among the predicted targets of the E2F1..5 motif. The categories are sorted by their enrichment, i.e. how much more frequent targets from this category are than expected by chance (first column) and only categories that are significantly enriched at a p-value of 0.05 (second column) are shown. The third and fourth columns in the table show the GO identifier and a description of the categories and these are again links to pages with more extensive information on the GO category. Finally, the user can interactively change the number of top categories shown using the drop-down menu or search for keywords.

Activity-expression correlation:

Gene	Promoter	Pearson	P-value	Plot
E2f1	chr2_-_154394864	0.69	1.3e-26	Click!
E2f2	chr4_+_135728600	0.68	1.7e-26	Click!
E2f3	chr13_-_30077931	0.63	7.5e-22	Click!
E2f4	chr8_+_107828491	0.53	1.8e-14	Click!
E2f5	chr3_+_14578783	-0.02	7.5e-01	Click!

Figure 4.16: Correlations between the E2F1..5 motif activity and mRNA expression profiles of TFs that kind bind to sites of the motif. The table shows the names of the associated TF genes, the IDs of the associated promoters of these genes, the Pearson correlation coefficient, the p-value for the correlation, and a link to a figure showing a scatter of the motif activity and mRNA expression levels across the samples (Suppl. Fig. 4.17) below.

Suppl. Fig. 4.18 shows the mRNA expression levels of the TF REST against its inferred motif activity, across the mouse GNF samples. The clear negative correlation strongly suggests that REST acts as a *repressor* of its targets, and this is indeed well-known to be the case.

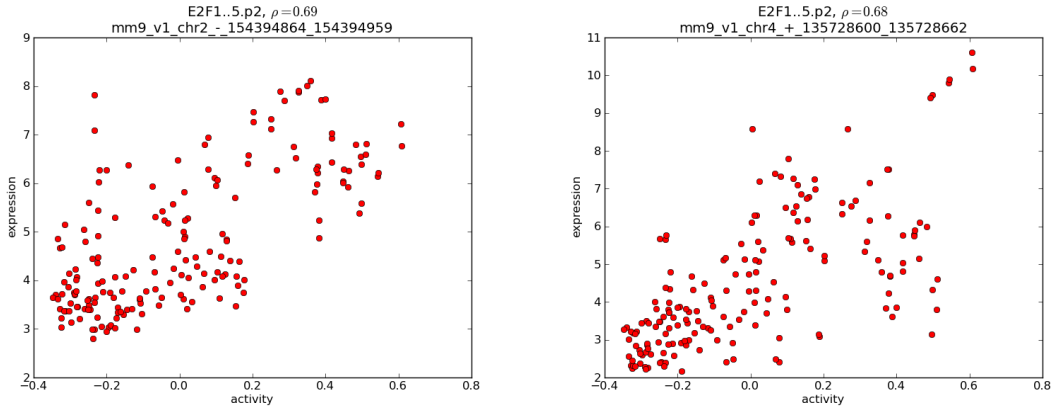


Figure 4.17: Example scatter plots showing the correlations between E2F1..5 motif activity and the mRNA expression of the E2F1 (left panel) and E2F2 (right panel) TFs, across the samples of the mouse GNF atlas. Each dot corresponds to one sample. The expression levels are shown on a logarithmic scale. At the top of the panel the Pearson correlation coefficient ρ and the ID of the promoter are shown.

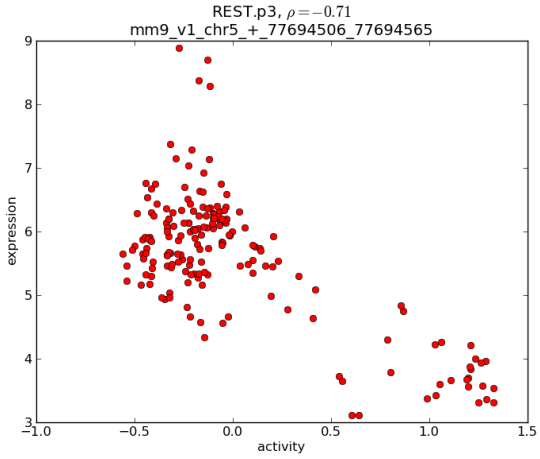


Figure 4.18: Scatter plots showing the correlation between REST motif activity and the mRNA expression of the REST TF, across the samples of the mouse GNF atlas. Each dot corresponds to one sample. The expression levels are shown on a logarithmic scale. At the top of the panel the Pearson correlation coefficient ρ and the ID of the promoter are shown.

Finally, one of our our aims is to understand the causal structure of the transcription regulatory network, and a first step in that direction are predictions of direct regulatory interactions between the motifs. For each motif, we check its list of predicted targets for promoters of TFs that are associated with other motifs. Using this we build a regulatory network where nodes correspond to motifs and a directed edge from motif m to motif m' occurs whenever a promoter of at least one of the TFs associated with motif m' is a predicted target of motif m. On the page with results of a given motif, a part of this regulatory network centered around the

motif in question is shown, i.e. all edges from or to the motif in question as well as edges between the direct neighbors of the motif. Supplementary Fig. 4.19 shows this network for the E2F1..5 motif. Note that a slider on the left-hand side of the network allows the user to vary a cut-off on the target score S_{pm}, i.e. showing only nodes and edges over the cut-off. In addition, placing the mouse pointer over a node brings up a pop-up with the z-value of the motif, and placing the mouse pointer on an edge will bring up a pop-up with the target score of the link.

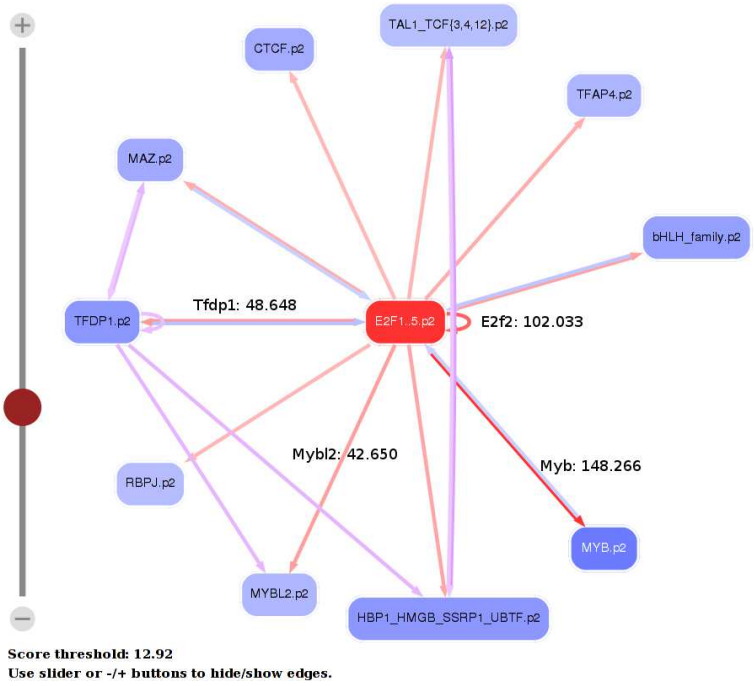


Figure 4.19: Predicted direct regulatory interactions between E2F and other motifs. Edges are drawn from motif m to m' whenever a promoter p, associated with motif m', is a predicted target of motif m, with a target score S_{pm} larger than a given cut-off c. In the web browser, the user can interactively change the cut-off c using the slider on the left of the figure. In this example the cut-off was set at 12.92. When the cursor is placed on an edge the target score S_{pm} is shown, and in this figure the target scores of the 4 most significant targets are shown. The intensity of the color of each motif corresponds to its z-score. Finally, for each motif (in this case E2F1..5) only the direct neighborhood in the network is shown, i.e. edges that are directly linked to E2F1..5, or that link between motifs that directly link to E2F1..5.

4.8 HNF1a activity in pancreas

Besides its well-known role in liver and kidney, ISMARA also predicts that HNF1a is one of the most active motifs in pancreas. Supplementary Fig. 4.20 shows the motifs with the most positive and most negative z-values in the two pancreas samples of the mouse GNF atlas.

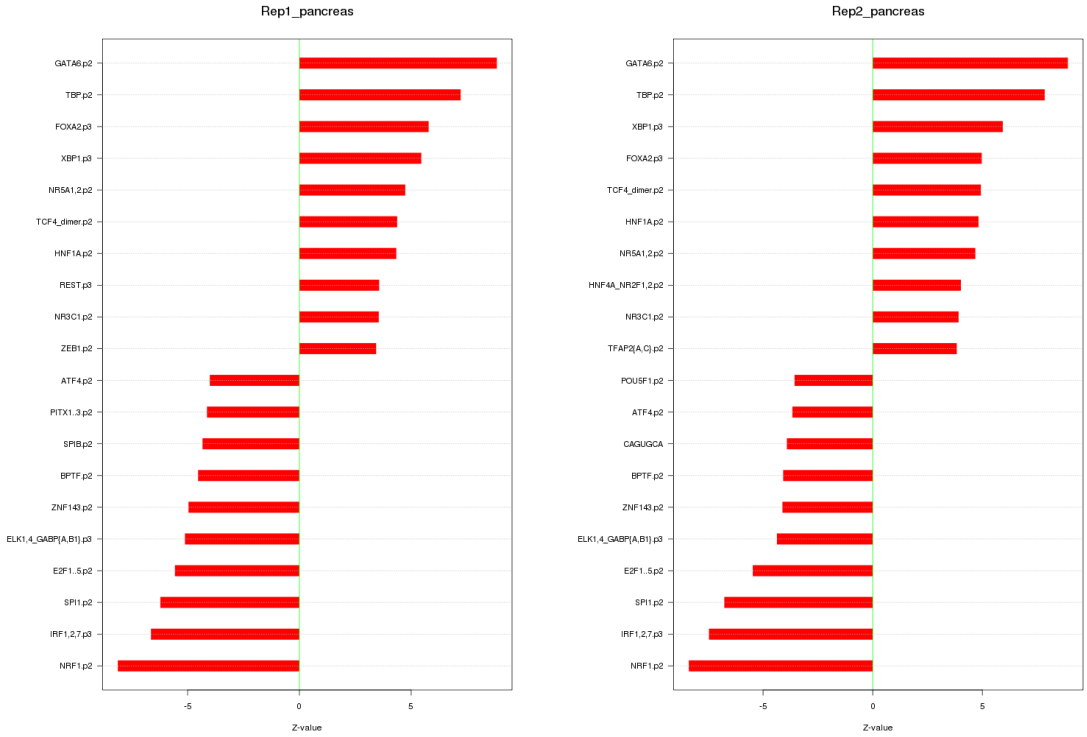


Figure 4.20: Motifs with the most positive and most negative z-values in two replicate pancreas samples from the mouse GNF atlas. Note that the HNF1a is the 7th and 6th most upregulated motif, respectively, in these samples.

4.9 Reproducibility of motif activities

The inferred motif activities depend both on our binding site predictions, and on the assumed simple linear relationship between predicted numbers of sites and mRNA expression. As explained in the main text, there are many reasons why such a 'cartoon' model is very unlikely to produce an accurate quantitative model of genome-wide expression profiles. As a consequence, one may wonder how robust the inferred motif activities are. However, as shown in Suppl. Figure 4.21, the motif activities inferred from the two replicates of the mouse GNF atlas are typically more reproducible across these replicates than the expression levels of individual promoters which are used to infer the motif activities. The reason for this is that the motif activity is inferred from the behavior of the hundreds to thousands of predicted targets of the motif. Thus, although at each individual promoter the expression is likely a complex function of the regulatory sites and the linear model is likely a poor approximation, these complications are effectively averaged out when inferring motif activities from the joint behavior of all targets.

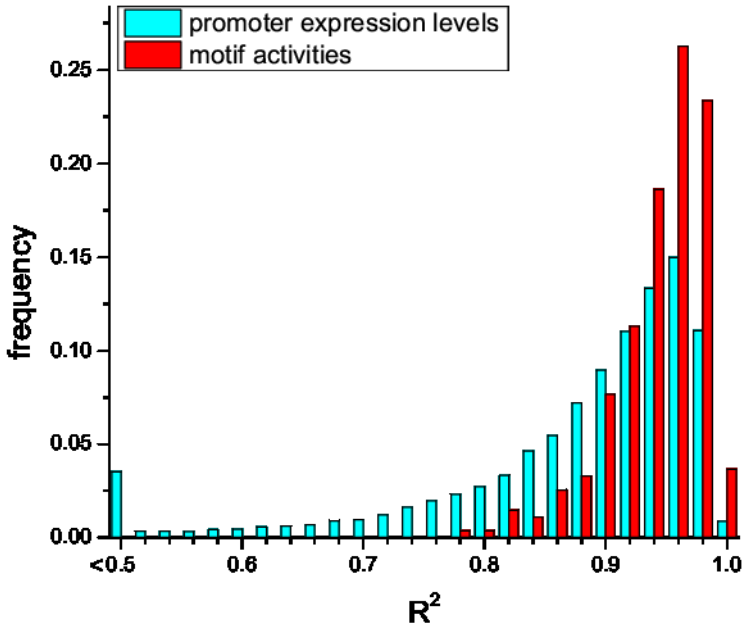


Figure 4.21: Reproducibility of the inferred motif activities and the expression profiles of promoters. For each motif, and each promoter, we calculated the Pearson correlation coefficient of the activity/expression profiles for the two replicates of the samples in the mouse GNF atlas. The figure shows the distribution of observed correlation coefficients for the motif activities (red) and promoter expression profiles (blue). The motif activities are generally considerably more reproducible than the expression profiles of the promoters from which they are inferred.

4.10 Motifs dis-regulated in tumor cells

To identify motifs whose motif activities are consistently dis-regulated in tumors, we first separate all samples s from the GNF and NCI-60 data sets into the set of tumor samples T and non-tumor samples N. Next, we use the replicate averaging described in section 4.5.8 to calculate, for each motif, an average activity $\langle \bar{A}^T \rangle$ in tumor samples, an associated error-bar $\delta \bar{A}^T$, an average activity in non-tumor samples $\langle \bar{A}^N \rangle$, and an error-bar $\delta \bar{A}^N$ associated with the average activity in non-tumor samples. From these, we calculate a z-value z_m for each motif m that quantifies the significance of the difference in the average activities in tumor and non-tumor samples. Tables 4.2 and 4.3 show the motifs with highest and lowest z-values, respectively. That is, these are the motifs most significantly dis-regulated in tumor cells.

Motif	z-values
bHLH_family.p2	2.398858
HIF1A.p2	2.230493
E2F1..5.p2	2.140652
ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2	2.071274
BPTF.p2	1.977484
NFY{A,B,C}.p2	1.920594
FOXD3.p2	1.915846
TFDP1.p2	1.901083
ELF1,2,4.p2	1.874818
ZNF143.p2	1.802732
ATF4.p2	1.786143
YY1.p2	1.735238
EHF.p2	1.718308
NRF1.p2	1.674024
ELK1,4_GABP{A,B1}.p3	1.667680
CCUUCAU (hsa-miR-205)	1.525379
PAX5.p2	1.500615
UCAAGUA (hsa-miR-26a, hsa-miR-26b, hsa-miR-1297, hsa-miR-4465)	1.404557
BACH2.p2	1.371868
GUAACAG (hsa-miR-194)	1.349047
HES1.p2	1.317505

Table 4.2: Motifs that are most consistently upregulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their z-value (shown in the second column).

4.11 XBP1 motif activity and mRNA expression

The XPB1 motif is the third most significant motif in the innate immune response time course in which HUVEC cells were treated with TNFα. The motif is upregulated during the time course. However, as shown in Suppl. Fig. 4.22, the mRNA expression of the XPB1 gene is almost constant across the time course, and not significantly correlated with the motif's activity. In fact, it has been established that XBP1's activity is regulated post-transcriptionally, i.e. through alternative splicing [55, 56].

4.12 Analysis of the ENCODE ChIP-seq data

To illustrate ISMARA's performance on ChIP-seq data we used data from the ENCODE project in which expression and 9 different chromatin modifications were measured across 8 different cell types [70]. Supple-

Motif	z-values
SMAD1..7,9.p2	-2.194113
HAND1,2.p2	-2.185943
TGIF1.p2	-2.117814
MAZ.p2	-2.076224
TFCP2.p2	-2.071225
KLF12.p2	-1.958392
GGCUCAG (hsa-miR-24)	-1.918863
FOX{D1,D2}.p2	-1.839199
TBX4,5.p2	-1.805228
FOXP3.p2	-1.740035
EVI1.p2	-1.701934
HBP1_HMGB_SSRP1_UBTF.p2	-1.688854
AAAGUGC (hsa-miR-17, hsa-miR-20a, hsa-miR-20b, hsa-miR-93, hsa-miR-106a, hsa-miR-106b, hsa-miR-519d)	-1.628037
GAGAUGA (hsa-miR-143, hsa-miR-4770)	-1.619611
HIC1.p2	-1.607936
NANOG{mouse}.p2	-1.576193
FEV.p2	-1.574951
MYOD1.p2	-1.565920
NR1H4.p2	-1.562673
POU1F1.p2	-1.556216
TCF4_dimer.p2	-1.536692
MYFfamily.p2	-1.514719
TAL1_TCF{3,4,12}.p2	-1.499900
POU5F1.p2	-1.480033
NR3C1.p2	-1.473553
HOX{A5,B5}.p2	-1.440485
STAT1,3.p3	-1.417964
GTF2A1,2.p2	-1.416557
RORA.p2	-1.391819
CAGCAGG (hsa-miR-214, hsa-miR-761, hsa-miR-3619-5p)	-1.356781
ETS1,2.p2	-1.337667
EN1,2.p2	-1.337051
AR.p2	-1.330996
RREB1.p2	-1.330444
CUCCCAA (hsa-miR-150)	-1.318296
CACAGUG (hsa-miR-128)	-1.318135
JUN.p2	-1.313498

Table 4.3: Motifs that are most consistently down-regulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their z-value (shown in the second column).

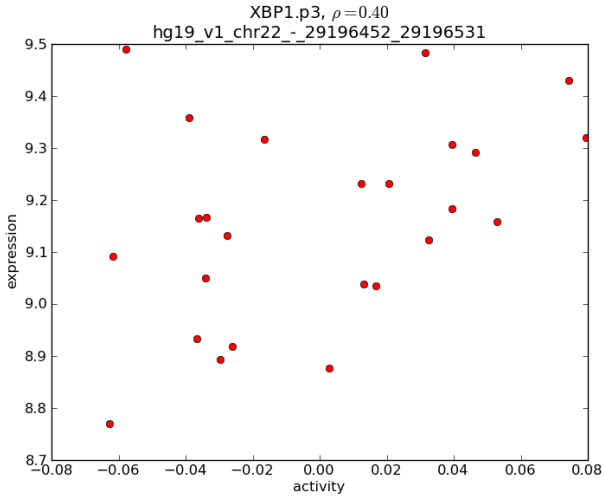


Figure 4.22: Scatter plot showing the correlation between the inferred activity of the XBP1 motif and the mRNA expression of the XBP1 gene for the innate immune response time course. The mRNA expression is shown on a logarithmic scale (base 2) along the vertical axis. Note the small range in expression variation.

Cell	Description
GM12878	B-lymphocyte, lymphoblastoid
HepG2	hepatocellular carcinoma
HMEC	mammary epithelial cells
HSMM	skeletal muscle myoblasts
Huvec	umbilical vein endothelial cells
K562	chronic myelogenous leukemia
NHEK	epidermal keratinocytes
NHLF	lung fibroblasts

Table 4.4: Human tissues and cell lines used as the source of experimental material in the ENCODE data sets for which we analyzed ChIP-seq data of chromatin marks. We used all available samples for which a consistent measurement platform was used.

mentary table 4.4 shows the list of cell types used together with their description and Suppl. table 4.5 shows a list of all the signals that were measured. For simplicity, we will refer to all 10 signals (which include expression and the binding of the CTCF transcription factor) as 'marks' in our description below.

We first ran ISMARA separately on the data sets for each of the 10 signals. For all the ChIP-seq data we thus modeled the occurrence of each of the marks at promoters in terms of the predicted TFBSs at the promoters. Supplementary table 4.6 lists all the data sets that were analyzed in this paper and shows, including references to the original publications, and lists for each data set the URL at which ISMARA's results for the corresponding data set can be found. Note that, for data sets 1, 2, and 5, there are also replicate averaged results available. These can be found by replacing 'ismara_report' at the end of the URL with 'averaged_report'.

Profiling	Platform
expression	Affymetrix HT Human Genome U133A Array
H3K4me3	Illumina Genome Analyzer II
H3K27me3	Illumina Genome Analyzer II
H3K27ac	Illumina Genome Analyzer II
H3K9ac	Illumina Genome Analyzer II
H3K36me3	Illumina Genome Analyzer II
H3K4me1	Illumina Genome Analyzer II
CTCF	Illumina Genome Analyzer II
H3K4me2	Illumina Genome Analyzer II
H4K20me1	Illumina Genome Analyzer II

Table 4.5: List of the signals (i.e. expression, histone modifications, and the binding of one TF) and corresponding measurement platforms from ENCODE data sets, that we used to demonstrate ISMARA's performance on ChIP-seq data sets. We used available BED and CEL files from the GSE26386 and GSE26312 GEO series.

Data Set	ISMARA URL
GNF SymAtlas, mouse [17]	ismara.unibas.ch/supp/dataset1/ismara_report
GNF SymAtlas + NCI-60 cancer cell lines, human [27, 28]	ismara.unibas.ch/supp/dataset2/ismara_report
Inflammatory response time course, HUVEC [42]	ismara.unibas.ch/supp/dataset3/ismara_report
Mucociliary differentiation, bronchial epithelial cells, human [57]	ismara.unibas.ch/supp/dataset4/ismara_report
Epithelial-Mesenchymal Transition, human [61]	ismara.unibas.ch/supp/dataset5/ismara_report
ENCODE cell lines, expression [70]	ismara.unibas.ch/supp/dataset6.1_ENCODE_expression/ismara_report
ENCODE cell lines, H3K4me3 [70]	ismara.unibas.ch/supp/dataset6.2_ENCODE_H3K4me3/ismara_report
ENCODE cell lines, H3K27me3 [70]	ismara.unibas.ch/supp/dataset6.3_ENCODE_H3K27me3/ismara_report
ENCODE cell lines, H3K27ac [70]	ismara.unibas.ch/supp/dataset6.4_ENCODE_H3K27ac/ismara_report
ENCODE cell lines, H3K9ac [70]	ismara.unibas.ch/supp/dataset6.5_ENCODE_H3K9ac/ismara_report
ENCODE cell lines, H3K36me3 [70]	ismara.unibas.ch/supp/dataset6.6_ENCODE_H3K36me3/ismara_report
ENCODE cell lines, H3K4me1 [70]	ismara.unibas.ch/supp/dataset6.7_ENCODE_H3K4me1/ismara_report
ENCODE cell lines, CTCF [70]	ismara.unibas.ch/supp/dataset6.8_ENCODE_CTCF/ismara_report
ENCODE cell lines, H3K4me2 [70]	ismara.unibas.ch/supp/dataset6.9_ENCODE_H3K4me2/ismara_report
ENCODE cell lines, H4K20me1 [70]	ismara.unibas.ch/supp/dataset6.10_ENCODE_H4K20me1/ismara_report

Table 4.6: URLs with the results of ISMARA's analyses of the data sets discussed in this paper.

4.12.1 PCA analysis

We first performed principal component analysis of the 10 marks across all promoters genome-wide, separately for each of the 8 cell types, as described in section 4.5.10. As shown in Suppl. Fig. 4.24, we find that the first principal component explains approximately 60% of the variation in each of the 8 cell types. In addition, the first principal component is almost identical in each of the cell types. This strongly suggests that this first principal component is a general feature of the distribution of chromatin marks. Moreover, the fact that this component aligns positively with expression and activity-associated chromatin marks, suggests that this first component reflects general promoter activity. We then pooled the data from all samples and performed principal component analysis on this complete data set, i.e. treating each promoter sample combination (p, s) as if it were a separate promoter. The resulting first principal component is shown in Fig. 6B of the main article.

Next, as described in section 4.5.10, we took the inferred motif activities for all marks and removed the component along the first principal component. That is, we removed the contribution to the motif activities that comes from the general 'promoter activity'. As an illustration, Suppl. Fig. 4.23 shows the inferred motif activities for 5 motifs (SNAI, IRF, HNF4a_NR2F1, TEAD1, and GATA6) both before (left panels) and after (right panels) the contribution from general promoter activity has been removed, for expression and the activation associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3. As the figure shows, before removal of the first PCA component, the activities for all marks are highly correlated, but this correlation disappears when the first PCA component is removed. This confirms that the highly correlated motif activities and the activation-associated chromatin marks is accounted for by the first PCA component that captures the relative chromatin mark levels associated with the general activity of a promoter. The remaining activities (right panels) thus provide a clearer insight in the specific role of a motif for specific marks across the cell-types. For example, for the SNAI motif the two acetylation marks are highly positive in HepG2 cells, whereas expression and H3K36me3 are clearly negative. Thus, promoters carrying SNAI sites tend to have higher histone acetylation levels than expected based on their general activity, and lower gene expression and H3K36me3 levels than expected based on the general activity.

As described in section 4.5.10, after removing the contribution of the first principal component to the motif activities, we re-calculated significance z-values z_m^i for each motif m and each mark i. In addition, we calculated a specificity s_m^i which measures the fraction of the overall that is associated with mark i. That is, a motif m will be highly specific for mark i if it has a high z-value z_m^i, and low z-values for all other marks. To identify motifs that are either most significant or highly specific for particular marks, we plotted scatter plots showing the significance and specificity for each motif (Suppl. Fig. 4.25). In each of the scatters we have indicated in red those motifs that had either very high significance or high specificity for the motif. Interestingly, we often find that the motifs with highest significance for a particular mark also have high specificity. For example, HNF1a is both most significant and most specific for H3K4me2 levels in promoters. Not surprisingly, the occurrence of CTCF motifs is the most significant determinant of the observed levels of bound CTCF.

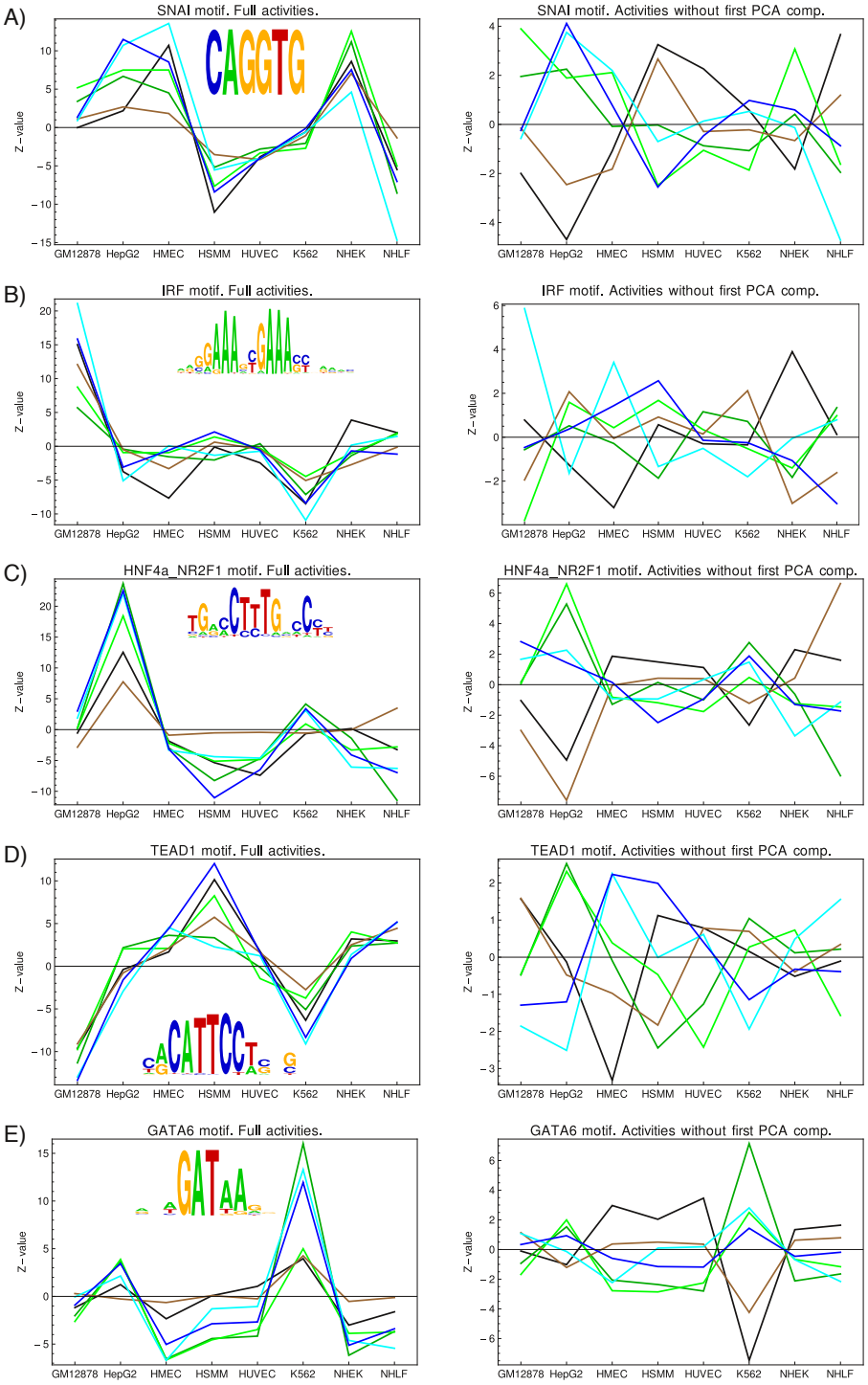


Figure 4.23: Inferred motif activities for 5 example motifs on the ENCODE ChIP-seq data sets measuring chromatin [70]. Each row (labeled A through E) shows the activities for explaining expression (black), H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown) levels, for one motif. The left panels show the motif activities as inferred from the original data the right panel the motif activities after the contribution along the first principal component has been subtracted. The names of the motifs are indicated above each panel and sequence logos are shown as insets. Note that the motif activities for the different marks go from highly correlated to essentially uncorrelated as the first principal component is removed.

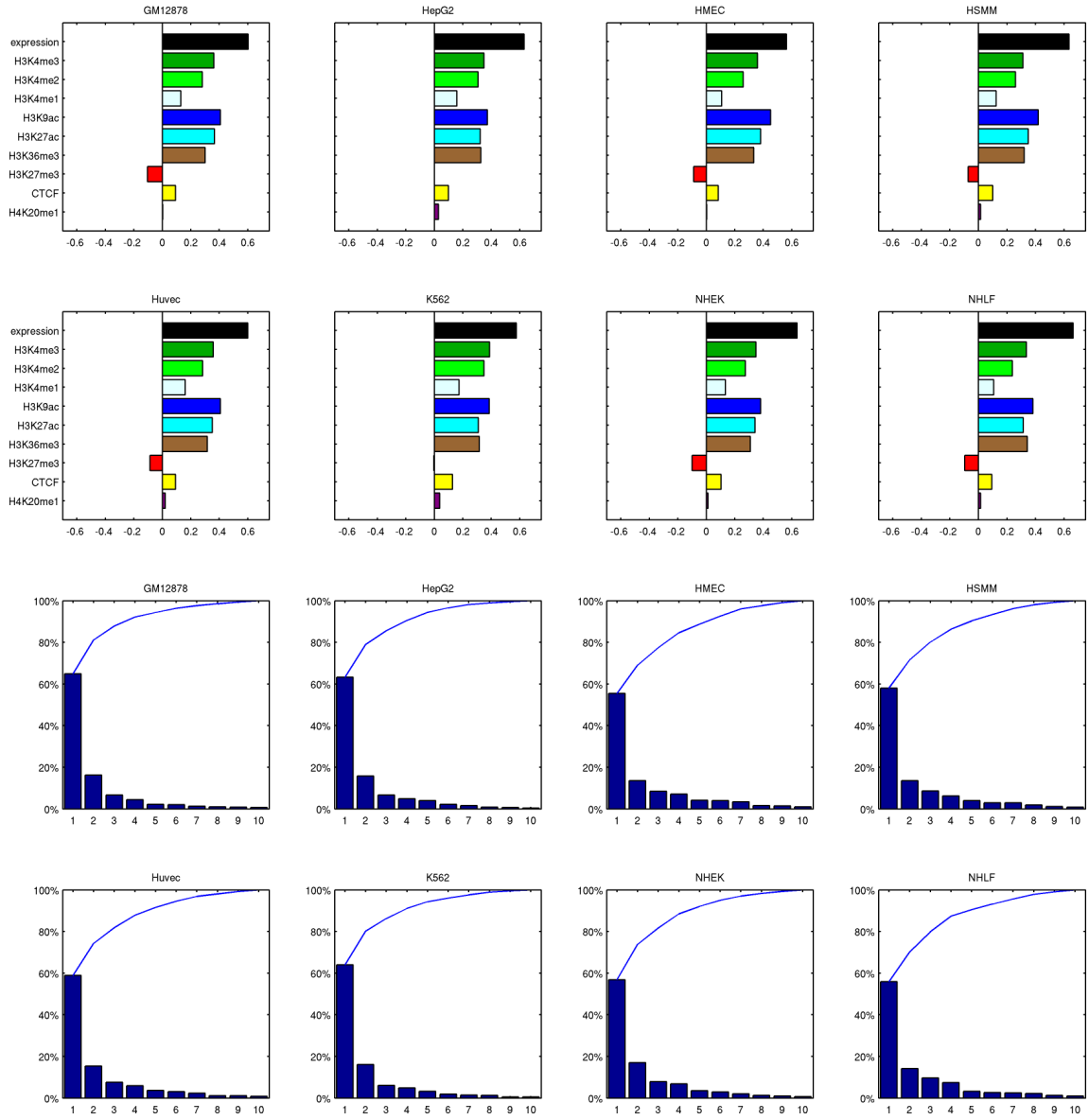


Figure 4.24: First principal component explaining the largest amount of chromatin mark and expression levels associated with each promoter, separately for each of the 8 cell types (top 8 panels). The bars indicate the relative contributions of expression and each of the chromatin marks to the first principal component. Note that the first principal component is virtually identical in each cell type. The bottom 8 panels show the fraction of the total variance explained by each subsequent principal component (bars) and the cumulative fraction of variance explained by consecutive components. Note that, for each cell type, close to 60% of the variance in expression and the 9 chromatin marks is explained by the first component.

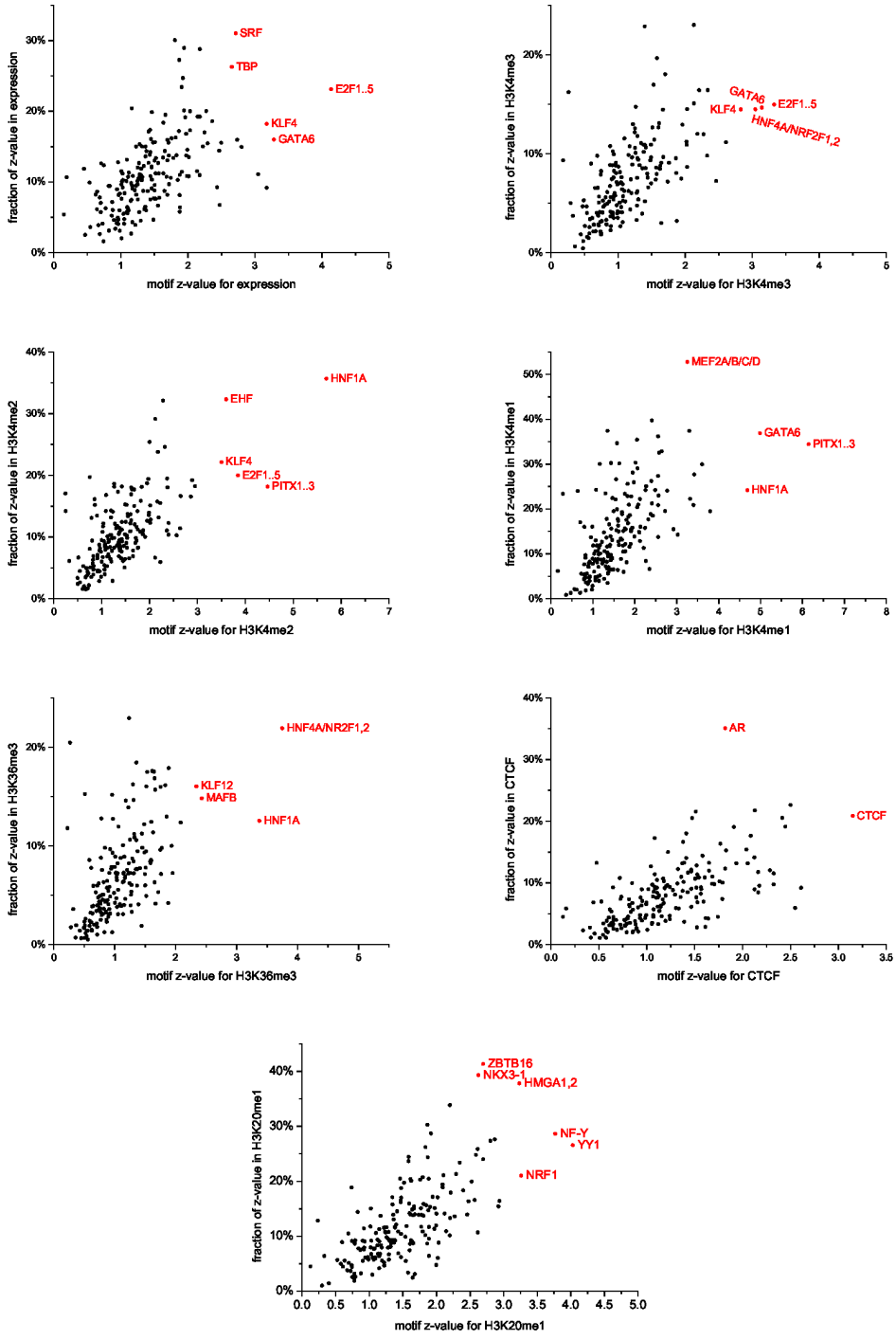


Figure 4.25: Significances and specificities of the motifs for explaining variations in different chromatin marks. Each panel corresponds to one mark (as indicated on the axes) and each dot corresponds to one motif. The significance of each motif is quantified by a z-value of the motif's activity for a given mark, after motif activities along the first principal component have been removed (see section 4.5.10). The specificity of a motif for a given mark is the fraction of all significance associated with a given mark (its z-value squared relative to the sum of all z-values squared, see section 4.5.10). the most significant and/or specific motifs for each mark are indicated in red.

Bibliography

[1] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.

[2] D.P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[3] M. R. Fabian, N. Sonenberg, and W. Filipowicz. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.*, 79:351–379, 2010.

[4] V. Matys, E. Fricke, R. Geffers, E. Gssling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378, Jan 2003.

[5] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004.

[6] Mikhail Pachkov, Ionas Erb, Nacho Molina, and Erik van Nimwegen. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*, 35(Database issue):D127–D131, Jan 2007.

[7] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263, Apr 2009.

[8] Erik van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6:S4, 2007.

[9] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, Jan 2009.

[10] Phil Arnold, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, 28(4):487–494, Feb 2012.

[11] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R R Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K.

Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P T Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schnbach, K. Sekiguchi, C. A M Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, F. A. N. T. O. M. Consortium, R. I. K. E. N. Genome Exploration Research Group, and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, Sep 2005.

[12] Piotr J Balwierz, Piero Carninci, Carsten O Daub, Jun Kawai, Yoshihide Hayashizaki, Werner Van Belle, Christian Beisel, and Erik van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol*, 10(7):R79, 2009.

[13] FANTOM Consortium and RIKEN Omics Science Center. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet*, 41(5):553–562, May 2009.

[14] Michiel de Hoon and Yoshihide Hayashizaki. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, 44(5):627–8, 630, 632, Apr 2008.

[15] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10096–100, 2000.

[16] Dat H Nguyen and Patrik D'haeseleer. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol Syst Biol*, 2:2006.0012, 2006.

[17] Jane E Lattin, Kate Schroder, Andrew I Su, John R Walker, Jie Zhang, Tim Wiltshire, Kaoru Saijo, Christopher K Glass, David A Hume, Stuart Kellie, and Matthew J Sweet. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res*, 4(1):5, 2008.

[18] N. B. La Thangue. DRTF1/E2F: an expanding family of heterodimeric transcription factors implicated in cell-cycle control. *Trends Biochem Sci*, 19(3):108–114, Mar 1994.

[19] J. M. Trimarchi and J. A. Lees. Sibling rivalry in the E2F family. *Nat. Rev. Mol. Cell Biol.*, 3:11–20, Jan 2002.

[20] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering.

STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–D416, Jan 2009.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[22] M. R. Campanero, M. Armstrong, and E. Flemington. Distinct cellular factors regulate the c-myb promoter through its E2F element. *Mol. Cell. Biol.*, 19(12):8442–8450, Dec 1999.

[23] M. S. Longworth, R. Wilson, and L. A. Laimins. HPV31 E7 facilitates replication by activating E2F2 transcription through its interaction with HDACs. *EMBO J.*, 24(10):1821–1830, May 2005.

[24] C. J. Kuo, P. B. Conley, C. L. Hsieh, U. Francke, and G. R. Crabtree. Molecular cloning, functional expression, and chromosomal localization of mouse hepatocyte nuclear factor 1. *Proc. Natl. Acad. Sci. U.S.A.*, 87:9838–9842, Dec 1990.

[25] M. S. Serfas and A. L. Tyner. HNF-1 alpha and HNF-1 beta expression in mouse intestinal crypts. *Am. J. Physiol.*, 265:G506–513, Sep 1993.

[26] M. Pontoglio, J. Barra, M. Hadchouel, A. Doyen, C. Kress, J. P. Bach, C. Babinet, and M. Yaniv. Hepatocyte nuclear factor 1 inactivation results in hepatic dysfunction, phenylketonuria, and renal Fanconi syndrome. *Cell*, 84:575–585, Feb 1996.

[27] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, Apr 2004.

[28] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–235, Mar 2000.

[29] G. L. Semenza. Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics. *Oncogene*, 29(5):625–634, Feb 2010.

[30] N. Meyer and L. Z. Penn. Reflecting on 25 years with MYC. *Nat. Rev. Cancer*, 8(12):976–990, Dec 2008.

[31] H. Z. Chen, S. Y. Tsai, and G. Leone. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer*, 9(11):785–797, Nov 2009.

[32] P. Gandellini, M. Folini, N. Longoni, M. Pennati, M. Binda, M. Colecchia, R. Salvioni, R. Supino, R. Moretti, P. Limonta, R. Valdagni, M. G. Daidone, and N. Zaffaroni. miR-205 Exerts tumor-suppressive functions in human prostate through down-regulation of protein kinase Cepsilon. *Cancer Res.*, 69(6):2287–2295, Mar 2009.

[33] S. Majid, A. A. Dar, S. Saini, S. Yamamura, H. Hirata, Y. Tanaka, G. Deng, and R. Dahiya. MicroRNA-205-directed transcriptional activation of tumor suppressor genes in prostate cancer. *Cancer*, 116(24):5637–5649, Dec 2010.

[34] A. A. Dar, S. Majid, D. de Semir, M. Nosrati, V. Bezrookove, and M. Kashani-Sabet. miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein. *J. Biol. Chem.*, 286(19):16606–16614, May 2011.

[35] H. Wu, S. Zhu, and Y. Y. Mo. Suppression of cell growth and invasion by miR-205 in breast cancer. *Cell Res.*, 19(4):439–448, Apr 2009.

[36] S. Liu, M. T. Tetzlaff, A. Liu, B. Liegl-Atzwanger, J. Guo, and X. Xu. Loss of microRNA-205 expression is associated with melanoma progression. *Lab. Invest.*, 92(7):1084–1096, Jul 2012.

[37] J. Kota, R. R. Chivukula, K. A. O'Donnell, E. A. Wentzel, C. L. Montgomery, H. W. Hwang, T. C. Chang, P. Vivekanandan, M. Torbenson, K. R. Clark, J. R. Mendell, and J. T. Mendell. Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell*, 137(6):1005–1017, Jun 2009.

[38] L. He, J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond. A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, Jun 2005.

[39] R. Chhabra, R. Dubey, and N. Saini. Cooperative and individualistic functions of the microRNAs in the miR-23a 27a 24-2 cluster and its implication in human diseases. *Mol. Cancer*, 9:232, 2010.

[40] K. H. To, S. Pajovic, B. L. Gallie, and B. L. Theriault. Regulation of p14ARF expression by miR-24: a potential mechanism compromising the p53 response during retinoblastoma development. *BMC Cancer*, 12:69, 2012.

[41] A. Lal, F. Navarro, C. A. Maher, L. E. Maliszewski, N. Yan, E. O'Day, D. Chowdhury, D. M. Dykxhoorn, P. Tsai, O. Hofmann, K. G. Becker, M. Gorospe, W. Hide, and J. Lieberman. miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. *Mol. Cell*, 35(5):610–625, Sep 2009.

[42] Youichiro Wada, Yoshihiro Ohta, Meng Xu, Shuichi Tsutsumi, Takashi Minami, Kenji Inoue, Daisuke Komura, Jun'ichi Kitakami, Nobuhiko Oshida, Argyris Papantonis, Akashi Izumi, Mika Kobayashi, Hiroko Meguro, Yasuharu Kanki, Imari Mimura, Kazuki Yamamoto, Chikage Mataki, Takao Hamakubo, Katsuhiko Shirahige, Hiroyuki Aburatani, Hiroshi Kimura, Tatsuhiko Kodama, Peter R Cook, and Sigeo Ihara. A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A*, 106(43):18357–18361, Oct 2009.

[43] Kenji Inoue, Mika Kobayashi, Kiichiro Yano, Mai Miura, Akashi Izumi, Chikage Mataki, Takeshi Doi, Takao Hamakubo, Patrick C. Reid, David A. Hume, Minoru Yoshida, William C. Aird, Tatsuhiko Kodama, and Takashi Minami. Histone deacetylase inhibitor reduces monocyte adhesion to endothelium through the suppression of vascular cell adhesion molecule-1 expression. *Arterioscler Thromb Vasc Biol*, 26(12):2652–2659, Dec 2006.

[44] Sybille Kempe, Hans Kestler, Andrea Lasar, and Thomas Wirth. NF-kappaB controls the global pro-inflammatory response in endothelial cells: evidence for the regulation of a pro-atherogenic program. *Nucleic Acids Res*, 33(16):5308–5319, 2005.

[45] H. Harada, E. Takahashi, S. Itoh, K. Harada, T. A. Hori, and T. Taniguchi. Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system. *Mol Cell Biol*, 14(2):1500–1509, Feb 1994.

[46] R. M. Ten, V. Blank, O. Le Bail, P. Kourilsky, and A. Isral. Two factors, IRF1 and KBF1/NF-kappa B, cooperate during induction of MHC class I gene expression by interferon alpha beta or Newcastle disease virus. *C R Acad Sci III*, 316(5):496–501, 1993.

[47] Gisline Martins and Kathryn Calame. Regulation and functions of Blimp-1 in T and B lymphocytes. *Annu Rev Immunol*, 26:133–169, 2008.

[48] Ulrike Seifert, Lukasz P. Bialy, Frdric Ebstein, Dawadschargal Bech-Otschir, Antje Voigt, Friederike Schrter, Timour Prozorovski, Nicole Lange, Janos Steffen, Melanie Rieger, Ulrike Kuckelkorn, Orhan Aktas, Peter-M. Kloetzel, and Elke Krger. Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell*, 142(4):613–624, Aug 2010.

[49] L. H. Glimcher. XBP1: the last two decades. *Ann Rheum Dis*, 69 Suppl 1:i67–i71, Jan 2010.

[50] Peter S. Gargalovic, Nima M. Gharavi, Michael J. Clark, Joanne Pagnon, Wen-Pin Yang, Aiqing He, Amy Truong, Tamar Baruch-Oren, Judith A. Berliner, Todd G. Kirchgessner, and Aldons J. Lusis. The unfolded protein response is an important regulator of inflammatory genes in endothelial cells. *Arterioscler Thromb Vasc Biol*, 26(11):2490–2496, Nov 2006.

[51] Mete Civelek, Elisabetta Manduchi, Rebecca J. Riley, Christian J Stoeckert, Jr, and Peter F. Davies. Chronic endoplasmic reticulum stress activates unfolded protein response in arterial endothelium in regions of susceptibility to atherosclerosis. *Circ Res*, 105(5):453–461, Aug 2009.

[52] Masanori Kitamura. Control of NF-κB and inflammation by the unfolded protein response. *Int Rev Immunol*, 30(1):4–15, Feb 2011.

[53] Arthur Kaser, Ann-Hwee Lee, Andre Franke, Jonathan N. Glickman, Sebastian Zeissig, Herbert Tilg, Edward E S. Nieuwenhuis, Darren E. Higgins, Stefan Schreiber, Laurie H. Glimcher, and Richard S. Blumberg. XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. *Cell*, 134(5):743–756, Sep 2008.

[54] Jingming Li, Joshua J. Wang, and Sarah X. Zhang. Preconditioning with endoplasmic reticulum stress mitigates retinal endothelial inflammation via activation of X-box binding protein 1. *J Biol Chem*, 286(6):4912–4921, Feb 2011.

[55] H. Yoshida, T. Matsui, A. Yamamoto, T. Okada, and K. Mori. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell*, 107(7):881–891, Dec 2001.

[56] M. Calfon, H. Zeng, F. Urano, J. H. Till, S. R. Hubbard, H. P. Harding, S. G. Clark, and D. Ron. IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature*, 415(6867):92–96, Jan 2002.

[57] Andrea J Ross, Lisa A Dailey, Luisa E Brighton, and Robert B Devlin. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am J Respir Cell Mol Biol*, 37(2):169–185, Aug 2007.

[58] E. Bonnafe, M. Touka, A. AitLounis, D. Baas, E. Barras, C. Ucla, A. Moreau, F. Flamant, R. Dubruille, P. Couble, J. Collignon, B. Durand, and W. Reith. The transcription factor RFX3 directs nodal cilium development and left-right asymmetry specification. *Mol Cell Biol*, 24(10):4417–4427, May 2004.

[59] Loubna El Zein, Aouatef Ait-Lounis, Laurette Morl, Jolle Thomas, Brigitte Chhin, Nathalie Spassky, Walter Reith, and Bndicte Durand. RFX3 governs growth and beating efficiency of motile cilia in mouse and controls the expression of genes involved in human ciliopathies. *J Cell Sci*, 122(Pt 17):3180–3189, Sep 2009.

[60] G. C. Horvath, M. K. Kistler, and W. S. Kistler. RFX2 is a candidate downstream amplifier of A-MYB regulation in mouse spermatogenesis. *BMC Dev. Biol.*, 9:63, 2009.

[61] Christina Scheel, Elinor Ng Eaton, Sophia Hsin-Jung Li, Christine L. Chaffer, Ferenc Reinhardt, Kong-Jie Kah, George Bell, Wenjun Guo, Jeffrey Rubin, Andrea L. Richardson, and Robert A. Weinberg. Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. *Cell*, 145(6):926–940, Jun 2011.

[62] Kornelia Polyak and Robert A. Weinberg. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer*, 9(4):265–273, Apr 2009.

[63] Mingxia Xiong, Lei Jiang, Yang Zhou, Wenjing Qiu, Li Fang, Rouyun Tan, Ping Wen, and Junwei Yang. The miR-200 family regulates TGF-1-induced renal tubular epithelial to mesenchymal transition through Smad pathway by targeting ZEB1 and ZEB2 expression. *Am J Physiol Renal Physiol*, 302(3):F369–F379, Feb 2012.

[64] Ulrike Burk, Jrg Schubert, Ulrich Wellner, Otto Schmalhofer, Elizabeth Vincan, Simone Spaderna, and Thomas Brabletz. A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *EMBO Rep*, 9(6):582–589, Jun 2008.

[65] Philip A. Gregory, Andrew G. Bert, Emily L. Paterson, Simon C. Barry, Anna Tsykin, Gelareh Farshid, Mathew A. Vadas, Yeesim Khew-Goodall, and Gregory J. Goodall. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol*, 10(5):593–601, May 2008.

[66] Karen M. Hajra, David Y-S. Chen, and Eric R. Fearon. The SLUG zinc-finger protein represses E-cadherin in breast cancer. *Cancer Res*, 62(6):1613–1618, Mar 2002.

[67] M. L. Grooteclaes and S. M. Frisch. Evidence for a function of CtBP in epithelial gene regulation and anoikis. *Oncogene*, 19(33):3823–3828, Aug 2000.

[68] Tang F, Zhang R, He Y, Zou M, Guo L, and et al. MicroRNA-125b Induces Metastasis by Targeting STARD13 in MCF-7 and MDA-MB-231 Breast Cancer Cells. *PLoS ONE*, 7(5):e35435, 2012.

[69] P. Arnold, A. Scholer, M. Pachkov, P. Balwierz, H. J?rgensen, M. B. Stadler, E. van Nimwegen, and D. Schubeler. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res.*, Sep 2012.

[70] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[71] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, 40(7):897–903, Jul 2008.

[72] R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 107(7):2926–2931, Feb 2010.

[73] S. C. Tippmann, R. Ivanek, D. Gaidatzis, A. Scholer, L. Hoerner, E. van Nimwegen, P. F. Stadler, M. B. Stadler, and D. Schubeler. Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol. Syst. Biol.*, 8:593, 2012.

[74] X. Dong, M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigo, E. Birney, and Z. Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, 13(9):R53, Sep 2012.

[75] N. D. Heintzman and B. Ren. Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.*, 19(6):541–549, Dec 2009.

[76] L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B. K. Lee, N. C. Sheffield, S. Graf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, 21(10):1757–1767, Oct 2011.

[77] B. Schuettengruber and G. Cavalli. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development*, 136(21):3531–3542, Nov 2009.

[78] L. W. Yuan and J. E. Gambee. Histone acetylation by p300 is involved in CREB-mediated transcription on chromatin. *Biochim. Biophys. Acta*, 1541(3):161–169, Dec 2001.

[79] K. Masternak, N. Peyraud, M. Krawczyk, E. Barras, and W. Reith. Chromatin remodeling and extragenic transcription at the MHC class II locus control region. *Nat. Immunol.*, 4(2):132–137, Feb 2003.

[80] Q. Gan, P. Thiebaud, N. Theze, L. Jin, G. Xu, P. Grant, and G. K. Owens. WD repeat-containing protein 5, a ubiquitously expressed histone methyltransferase adaptor protein, regulates smooth muscle cell-selective gene activation through interaction with pituitary homeobox 2. *J. Biol. Chem.*, 286(24):21853–21864, Jun 2011.

[81] K. O. Kizer, H. P. Phatnani, Y. Shibata, H. Hall, A. L. Greenleaf, and B. D. Strahl. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Mol. Cell. Biol.*, 25(8):3305–3316, Apr 2005.

[82] W. Yuan, J. Xie, C. Long, H. Erdjument-Bromage, X. Ding, Y. Zheng, P. Tempst, S. Chen, B. Zhu, and D. Reinberg. Heterogeneous nuclear ribonucleoprotein L Is a subunit of human KMT3a/Set2 complex required for H3 Lys-36 trimethylation activity in vivo. *J. Biol. Chem.*, 284(23):15701–15707, Jun 2009.

[83] Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178):535–540, Jan 2008.

[84] Xin He, Md Abul Hassan Samee, Charles Blatti, and Saurabh Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol*, 6(9), Sep 2010.

[85] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, Aug 2006.

[86] Masaki Ieda, Ji-Dong Fu, Paul Delgado-Olguin, Vasanth Vedantham, Yohei Hayashi, Benoit G Bruneau, and Deepak Srivastava. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, 142(3):375–386, Aug 2010.

[87] H. D. Kim, T. Shay, E. K. O'Shea, and A. Regev. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, 325(5939):429–432, Jul 2009.

[88] Jianhua Ruan. A top-performing algorithm for the DREAM3 gene expression prediction challenge. *PLoS One*, 5(2):e8944, 2010.

[89] K. M. Summers, S. Raza, E. van Nimwegen, T. C. Freeman, and D. A. Hume. Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome. *Eur. J. Hum. Genet.*, 18(11):1209–1215, Nov 2010.

[90] Nicola Aceto, Nina Sausgruber, Heike Brinkhaus, Dimos Gaidatzis, Georg Martiny-Baron, Giovanni Mazzarol, Stefano Confalonieri, Micaela Quarto, Guang Hu, Piotr J. Balwierz, Mikhail Pachkov, Stephen J. Elledge, Erik van Nimwegen, Michael B. Stadler, and Mohamed Bentires-Alj. Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop. *Nat Med*, 18(4):529–537, 2012.

[91] Joaqun Prez-Schindler, Serge Summermatter, Silvia Salatino, Francesco Zorzato, Markus Beer, Piotr J. Balwierz, Erik van Nimwegen, Jrme N. Feige, Johan Auwerx, and Christoph Handschin. The Corepressor NCoR1 Antagonizes PGC-1α and ERRα in the Regulation of Skeletal Muscle Function and Oxidative Metabolism. *Mol Cell Biol*, Oct 2012.

[92] Erik Arner, Niklas Mejhert, Agn Kulyt, Piotr J. Balwierz, Mikhail Pachkov, Mireille Cormont, Silvia Lorente-Cebrin, Anna Ehrlund, Jurga Laurencikiene, Per Hedn, Karin Dahlman-Wright, Jean-Franois Tanti, Yoshihide Hayashizaki, Mikael Rydn, Ingrid Dahlman, Erik van Nimwegen, Carsten O. Daub, and Peter Arner. Adipose tissue microRNAs as regulators of CCL2 production in human obesity. *Diabetes*, 61(8):1986–1993, Aug 2012.

[93] R. Hasegawa, Y. Tomaru, M. de Hoon, H. Suzuki, Y. Hayashizaki, and J. W. Shin. Identification of ZNF395 as a novel modulator of adipogenesis. *Exp. Cell Res.*, Nov 2012.

[94] Q. Cui, Z. Yu, E.O. Purisima, and E. Wang. Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.*, 2:46, 2006.

[95] E. Hornstein and N. Shomron. Canalization of development by microRNAs. *Nat. Genet.*, 38 Suppl:S20–S24, 2006.

[96] Y. Zhou, J. Ferguson, J.T. Chang, and Y. Kluger. Inter- and intra-combinatorial regulation by transcription factors and microRNAs. *BMC Genomics*, 8:396, 2007.

[97] Denis C Bauer, Fabian A Buske, and Timothy L Bailey. Dual-functioning transcription factors in the developmental gene network of Drosophila melanogaster. *BMC Bioinformatics*, 11(1):366, Jul 2010.

[98] Martha L Bulyk. DNA microarray technologies for measuring protein-DNA interactions. *Curr Opin Biotechnol*, 17(4):422–430, Aug 2006.

[99] Nathaniel D Heintzman, Gary C Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, Keith A Ching, Jessica E Antosiewicz-Bourget, Hui Liu, Xinmin Zhang, Roland D Green, Victor V Lobanenkov, Ron Stewart, James A Thomson, Gregory E Crawford, Manolis Kellis, and Bing Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009.

[100] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, Dec 2011.

[101] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan 2008.

[102] C. Notredame, D. Higgins, and J. Heringa. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[103] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrst, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Taber, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. GRimmend, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, 38(6):626–635, 2006.

[104] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, and Carninci P. Cage: cap analysis of gene expression. *Nat Methods*, 3(3):211–22, Mar 2006. PMID: 16489339.

[105] W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12:656–664, 2002.

[106] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The ucsc genome browser database. *Nucl. Acids Res.*, 31(1):51–54, 2003.

[107] Elodie Portales-Casamar, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, and Albin Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 38:D105D110, 2010.

[108] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, J. S. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Research*, 36:D281–D288, 2008.

[109] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[110] Rahul Siddharthan, Eric D Siggia, and Erik van Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, Dec 2005.

[111] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, M. Airey, A. Anagnostopoulos, R. Babiuk, R. Baldarelli, J. Beal, S. Bello, N. Butler, J. Campbell, L. Corbani, S. Giannatto, H. Dene, M. Dolan, H. Drabkin, K. Forthofer, M. Knowlton, J. Lewis, M. McAndrews-Hill, S. McClatchy, D. Miers, L. Ni, H. Onda, J. E. Ormsby, J. Recla, D. Reed, B. Richards-Smith, R. Shaw, R. Sinclair, D. Sitnikov, C. Smith, L. Washburn, and Y. Zhu. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, 40(Database issue):D881–886, Jan 2012.

[112] Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C Hardison, David Haussler, and Webb Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, Jan 2003.

[113] Nacho Molina and Erik van Nimwegen. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res*, 18(1):148–160, Jan 2008.

[114] Webb Miller, Kate Rosenbloom, Ross C Hardison, Minmei Hou, James Taylor, Brian Raney, Richard Burhans, David C King, Robert Baertsch, Daniel Blankenberg, Sergei L Kosakovsky Pond, Anton Nekrutenko, Belinda Giardine, Robert S Harris, Svitlana Tyekucheva, Mark Diekhans, Thomas H Pringle, William J Murphy, Arthur Lesk, George M Weinstock, Kerstin Lindblad-Toh, Richard A Gibbs, Eric S Lander, Adam Siepel, David Haussler, and W. James Kent. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, 17(12):1797–1808, Dec 2007.

[115] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[116] Benilton S Carvalho and Rafael A. Irizarry. A Framework for Oligonucleotide Microarray Preprocessing. *Bioinformatics*, 2010.

[117] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99:909, 2004.

[118] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[119] C. Fraley and A. E. Raftery. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, (2006, revised in 2009).

[120] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, Dec 2004.

[121] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 1:25–29, May 2000.

[122] M. Pachkov, P. J. Balwierz, P. Arnold, E. Ozonov, and E. van Nimwegen. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, Nov 2012.

Chapter 5

Discussion and Future Work

In previous studies [1–5], it has been shown that there are several mechanisms for recruiting chromatin marks to genomic loci in mammals. One branch of research addresses the idea that TFs recruit chromatin modifications to their sites of action. Correlating the occurrence of DNA binding sites with the chromatin mark measured in one cell line or tissue has often been used to associate chromatin modifications with TFBSs []. We wanted to take the idea of TFs recruiting chromatin modifications a step further and decided to not simply correlate the absolute chromatin mark levels with TFBSs, but rather model the relative changes of chromatin levels, for a given mark across different cell stages, as a linear function of the expected number of predicted TFBSs in regulatory regions such as promoters, enriched chromatin mark regions, or cis regulatory modules. We named our algorithm Epi-MARA to point out the similarity to MARA [6], an algorithm that models changes in transcription in terms of predicted TFBSs.

We applied Epi-MARA to an in-vitro mouse neuronal differentiation system [7, 8]. Starting with embryonic stem cells (ES stage), they can be differentiated through neuronal progenitor cells (NP stage) into terminal neurons (TN stage).
First, we focused on the dynamics at promoters of the repressive chromatin modification H3K27me3, set by the Polycomb system [9]. Among the top predicted motifs that contributed most to explaining the H3K27me3 dynamics at promoters was the motif bound by the TF REST. Epi-MARA predicted that REST targets gain H3K27me3 when going from ES to NP and lose H3K27me3 when going from NP to TN (see figure 2.2(b) in chapter 2).
Secondly, we extended our investigation genome-wide to all H3K27me3 enriched regions. Based on their CpG content, the H3K27me3 enriched regions could be separated into two classes corresponding to either high-CpG or low-CpG regions. Running Epi-MARA on all H3K27me3 enriched regions, while taking into account the different base composition for low-CpG and high-CpG regions when predicting TFBS, revealed that REST recruits H3K27me3 at the NP stage at the high-CpG regions (as in the promoter case), whereas at the low-CpG regions it actually depletes the chromatin mark (see figure 2.3(b) in chapter 2). As the REST motif is only bound by a single TF and as RESTko cells can be created for all three time points in the neuronal differentiation systems, we decided to perform extensive tests to validate Epi-MARA's predictions for REST: **1.** By comparing H3K27me3 levels between wildtype and the RESTko for H3K27me3 REST targets (that is to say H3K27me3 enriched regions that overlap a REST ChIP-seq peak), we could show that in ES, there is little effect of REST on the H3K27me3 levels in neither high-CpG nor low-CpG regions. This is in stark contrast to the situation in the NP stage. There, high-CpG REST targets clearly have more H3K27me3 in wildtype, whereas low-CpG REST targets clearly have less H3K27me3 in wildtype. **2.** By looking at the fold-change of H3K27me3 levels across ES, NP, and TN, one clearly sees that the high-CpG and low-CpG regions behave in opposite fashion, with H3K27me3 REST targets showing even greater fold-change, in

agreement with Epi-MARA's prediction for REST. **3.** To show that the REST binding site is itself necessary to recruit H3K27me3 in high-CpG regions, we inserted an approximately 2kb long sequence (once containing the original REST binding site, once containing no REST binding site), belonging to a high-CpG region which is a REST target, into a H3K27me3 free region in the mouse genome. By comparing the H3K27me3 levels with or without the REST binding site to their endogenous levels in NP, we confirmed that H3K27me3 levels are significantly higher when the REST site is present.

I will now discuss the main findings of my PhD project and give an outlook of where future work could be taken up.

5.0.2 Biological relevance

Predicting and validating possible TFs involved in recruiting/depleting of chromatin marks at different cell stages, even though important, is only one of several steps towards understanding the function and effects of chromatin marks on gene expression patterns. In my opinion, one of the most striking biological findings from our work is that a DNA fragment of only 2kb length, which belongs to a high-CpG H3K27me3 region and contains a REST binding site, is enough to recruit H3K27me3. Inserting the same fragment with a deleted REST binding site leads to significantly lower increase of H3K27me3 levels. The fact that the H3K27me3 levels at the insert are (sometimes clearly) lower than at the endogenous region shows that the regulation of H3K27me3 levels is a complex process, that is to say that the local context encompassing the fragment is also of particular importance. It could contain other chromatin modifications and other binding sites for chromatin associated TFs (like in the case of PREs in fly [10]). It is not unlikely that there are even different mechanisms at work at the same time. It remains to be seen if REST is one of those factors that will eventually be identified as occuring in mammalian PRE-like regions.

Another interesting observation is that high-CpG and low-CpG regions show quite different behavior when comparing the fold-changes of H3K27me3 levels across the tree time points (see figure 2.3 in chapter 2). That is to say, the recruitment/depletion of H3K27me3 strongly depends on the stage of the cell. This is especially true in the case of REST, for which we predicted and experimentally confirmed that the H3K27me3 recruitment/depletion ability is much higher in the NP stage than in the ES stage. This immediately raises two questions:

- How is this stage dependent regulation achieved? and
- What is the reason for this temporal change in H3K27me3 at H3K27me3 REST targets?

It is very likely that the cell-stage dependence on H3K27me3 levels is achieved by cell-type specific components. This might be a set of TFs that are NP specific, and which interact with the local sequence or chromatin context of a targeted region. This could then lead to conformational changes to the protein (TF?) that interacts with Polycomb. Especially in the case of REST, I could imagine that one combination of TFs binding to high-CpG regions increases REST's ability to interact with Polycomb, whereas another combination of TFs binding to low-CpG regions decreases REST's ability to interact with Polycomb. I want to point out that there is no experimental proof that REST interacts directly with Polycomb!

Inferring the biological purpose of REST recruiting H3K27me3 in high-CpG regions and depleting H3K27me3 in low-CpG regions at the NP stage is not an easy task. One of the roles of REST is, as mentioned before, to repress neuronal genes in non-neuronal tissues. As REST protein levels have already decreased in the NP stage, one could imagine that REST recruits Polycomb to high-CpG targets in (or before) the NP stage to make sure that some neuronal genes that must not be expressed at that stage are kept shut off. By performing a gene ontology analysis [11, 12] on all H3K27me3 regions that are both associated with a gene (proximal

H3K27me3 regions) and also show a significant change in H3K27me3 levels when comparing wild-type to knock-out yields, not surprisingly, we see association with a lot of neuron-related categories. By looking closer at the categories, many of them are involved in the shape of the cell (synapses, ...). Therefore, a speculative hypothesis would be that H3K27me3 might prevent the cell from changing its shape too early.
On the other hand, low-CpG REST targets are depleted in H3K27me3 at the NP stage. As these low-CpG REST targets are mostly distal, it is not straightforward to see how to associate the region with a gene. It might be that some essential genes for the neuronal progenitor stage controlled by distal cis regulatory modules must be depleted of H3K27me3 for their proper function at this stage. Alternatively, the dynamics in H3K27me3 in these distal regions could be due to changes in the chromosomal structure.

5.0.3 Future work

Epi-MARA strongly depends on the predicted TFBSs. We showed that substituting predicted TFBSs for a TF by its real binding data, increased the significance of that motif by a factor of 2. Unfortunately, creating ChIP-seq data sets for more than a handful of factors is still unfeasible. Another way of improving the predicted TFBSs is by only including the regulatory regions that are accessible for TFs at the different time points or tissues. Using, for example, DNase-seq [13] one can assign to each region an accessibility score [14]. This procedure should reduce the number of falsely predicted sites by a fair amount [14].

Even though it is well established that H3K27me3 is a silencing mark [1, 2, 15, 16], that is to say associated with closed chromatin, it is still not known to what extend it contributes to this. One way of further investigating that in our neuronal differentiation system would be to predict TFBSs for REST genome-wide. Then, one would calculate how much binding you should obtain based on the DNA sequence itself, and compare this to how much binding there really is based on the ChIP-seq measurements. By looking at the levels of H3K27me3 (and different chromatin marks like H3K4...), one could get an idea of how much these chromatin marks are closing or opening the DNA.

The ultimate goal is to predict the regions in the genome enriched in a certain chromatin mark, based only on sequence composition. A first attempt was made by Leonie Ringrose et al. in trying to predict PREs in fly [10]. Basically, the idea was to use experimentally validated PREs and assign a score to the occurrence of clusters of two or three PRE associated motifs, namely GAF, PHO, and ZESTE. After training, all PREs in the training data set could be identified. Furthermore, when scanning the entire fly genome, about 330 PREs were predicted, containing all validated ones. One possibility to tackle this problem in mammals would be by combining Epi-MARA and MotEvo. In a first step, one uses Epi-MARA to predict key TFs involved in the recruitment/depletion of the chromatin mark of interest. In a second step, one slides a short ($\approx$ 2kb) window across the genome and calculates the score Z

$$Z = \frac{P(S|\omega, \text{bg})}{P(S|\text{bg})}, \tag{5.1}$$

which is the ratio of the probability of sequence S in this window, given the set of WMs ω corresponding to Epi-MARA's top predictions, and the background model bg, and the probability of sequence S given only the background model. Windows with a high Z are then predicted as enriched in the chromatin marks. Even though this approach sounds compelling, there could be several issues.
Experiments show that binding of PRC1 and PRC2 is spread across many kilobases of mammalian developmental genes [17–19]. If the chromatin mark is recruited by a set of TFs whose TFBSs are far apart, the sliding window approach might have problems in picking up these clusters of TFBSs. On the other hand, our analysis on the H3K27me3 enriched regions provide some evidence that the regions where such clusters of TFBSs occur do not need to be widespread; most of the H3K27me3 clusters are shorter than 4kb. As we have showed in the case of REST, recruitment of H3K27me3 is also dependent on the current stage of the cell. By

just looking at the raw DNA sequence, it is not possible to predict these transient increases. And we must not forget that it has been shown that there are also other mechanisms (small nc-RNA, ...) [1–5] that could recruit chromatin marks.

As a closing remark, I would like to say that in my opinion, the successful experimental validation of one of Epi-MARA's top predictions, that is REST recruiting H3K27me3 at high-CpG regions and depleting H3K27me3 in low-CpG regions, showed that using Epi-MARA to model chromatin dynamics is a promising approach that could take us a step further to understanding how chromatin marks are targeted to their sites of action, and what the detailed effects on the gene expression patterns and/or chromatin structure are.

Bibliography

[1] C. David Allis, Thomas Jenuwein, and Danny Reinberg. *Epigenetics*. Cold Spring Harbor Laboratory Press, 2007.

[2] Fabio Mohn. *Epigenome plasticity during cellular differentiation*. PhD thesis, FMI, 2009.

[3] Jeffrey A. Simon and Robert E. Kingston. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Biol*, 10:697–708, 2009.

[4] Aline V Probst, Elaine Dunleavy, and Genevieve Almouzni. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol*, 10:192–206, 2009.

[5] James Flanagan and Laurence Wild. An epigenetic role for noncoding rnas and intragenic dna methylation. *Genome Biology*, 8(6):307, 2007.

[6] The FANTOM Consortium and Riken Omics Science Center. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, 41:553–562, 2009.

[7] Miriam Bibel, Jens Richter, Katrin Schrenk, Kerry Lee Tucker, Volker Staiger, Martin Korte, Magdalena Goetz, and Yves-Alain Barde. Differentiation of mouse embryonic stem cells into a defined neuronal lineage. *Nature Neuroscience*, 7:1003–1009, 2004.

[8] Miriam Bibel, Jens Richter, Emmanuel Lacroix, and Yves-Alain Barde. Generation of a defined and uniform population of cns progenitors and neurons from mouse embryonic stem cells. *Nat. Protocols*, 2(5):1034, 1043.

[9] Raphael Margueron and Danny Reinberg. The polycomb complex prc2 and its mark in life. *Nature*, 469:343 – 349, 2011.

[10] Leonie Ringrose, Marc Rehmsmeier, Jean-Maurice Dura, and Renato Paro. Genome-wide prediction of polycomb/trithorax response elements in drosophila melanogaster. *Developmental Cell*, 5(5):759 – 771, 2003.

[11] Huang DW, Sherman BT, and Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protoc.*, 4:44–57, 2009.

[12] Huang DW, Sherman BT, and Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37:1–13, 2009.

[13] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-resolution mapping andcharacterization of open chromatin across the genome. *Cell*, 132(2):311 – 322, 2008.

[14] Roger Pique-Regi, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, Yoav Gilad, and Jonathan K. Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Res.*, 21:447–455, 2011.

[15] Valerio Orlando. Polycomb, epigenomes, and control of cell identity. *Cell*, 112(5):599 – 606, 2003.

[16] Stuart S. Levine, Ian F.G. King, and Robert E. Kingston. Division of labor in polycomb group repression. *Trends in Biochemical Sciences*, 29(9):478 – 485, 2004.

[17] Laurie A. Boyer, Kathrin Plath, Julia Zeitlinger, Tobias Brambrink, Lea A. Medeiros, Tong Ihn Lee, Stuart S. Levine, Marius Wernig, Adriana Tajonar, Mridula K. Ray, George W. Bell, Arie P. Otte, Miguel Vidal, David K. Gifford, Richard A. Young, and Rudolf Jaenisch. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441:349, 353.

[18] Tong Ihn Lee, Richard G. Jenner, Laurie A. Boyer, Matthew G. Guenther, Stuart S. Levine, Roshan M. Kumar, Brett Chevalier, Sarah E. Johnstone, Megan F. Cole, Kyo ichi Isono, Haruhiko Koseki, Takuya Fuchikami, Kuniya Abe, Heather L. Murray, Jacob P. Zucker, Bingbing Yuan, George W. Bell, Elizabeth Herbolsheimer, Nancy M. Hannett, Kaiming Sun, Duncan T. Odom, Arie P. Otte, Thomas L. Volkert, David P. Bartel, Douglas A. Melton, David K. Gifford, Rudolf Jaenisch, and Richard A. Young. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301 – 313, 2006.

[19] Manching Ku, Richard P. Koche, Esther Rheinbay, Eric M. Mendenhall, Mitsuhiro Endoh, Tarjei S. Mikkelsen, Aviva Presser, Chad Nusbaum, Xiaohui Xie, Andrew S. Chi, Mazhar Adli, Simon Kasif, Leon M. Ptaszek, Chad A. Cowan, Eric S. Lander, Haruhiko Koseki, and Bradley E. Bernstein. Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. *PLoS Genet*, 4(10):e1000242, 10 2008.